SURVEY METHODOLOGY

Catalogue No. 12-001-XPB

A JOURNAL PUBLISHED BY STATISTICS CANADA

DECEMBER 2001

•

VOLUME 27

•

NUMBER 2

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A JOURNAL PUBLISHED BY STATISTICS CANADA

DECEMBER 2001 • VOLUME 27 • NUMBER 2

Published by authority of the Minister
responsible for Statistics Canada

February 2002

Catalogue no. 12-001-XPB

Frequency: Semi-annual

ISSN 0714-0045

Ottawa

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

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

EDITORIAL POLICY

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

Submission of Manuscripts

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

Subscription Rates

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

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Volume 27, Number 2, December 2001

CONTENTS

Survey Methodology, December 2001
Vol. 27, No. 2, pp. 119-120
Statistics Canada

In This Issue

This issue of *Survey Methodology* contains papers on a variety of topics touching on coverage issues, nonresponse, imputation, survey designs, survey weighting and analysis of data from complex surveys.

In the first paper of this issue, Blenk and Stasny develop a weighting adjustment in order to reduce the coverage bias in telephone surveys while controlling the increase in variance due to weighting. The weighting adjustment is applied to *transient* households, which are households moving in and out of the telephone population during the year. It is assumed that the transient telephone population is representative of the non-telephone population. The weighting adjustment proposed is based on propensity scores for transience obtained using a logistic regression model. The proposed method and several alternatives are compared using data collected from a survey of distressed and non-distressed regions of Kentucky, Ohio, and West Virginia.

Mariano and Kadane use the information on the number of calls in a telephone survey as an indicator of how difficult an intended respondent is to reach. This permits a probabilistic division of the nonrespondents into those who will always refuse to respond and those who were not available to respond in a model of the nonresponse. It also permits an evaluation of whether the nonresponse is ignorable for inference about the dependent variable by incorporating the information on the number of calls into the model. These ideas are implemented on data from a survey in Metropolitan Toronto of attitudes toward smoking in the workplace. The results reveal that the nonresponse is not ignorable and those who do not respond are twice as likely to favor unrestricted smoking in the workplace as are those who do.

In his paper, Hidiroglou unifies the nested and non-nested cases found in the double sampling theory. The nested case, also known as two-phase sampling, corresponds to the traditional case in which a first-phase sample is initially taken so that additional information may be collected. This is followed by a second-phase sample taken within the first one, which contains the variables of interest. The non-nested case reflects a situation in which both samples are selected independently from the same frame or possibly from different frames. Using the generalized difference, an estimator is proposed for both cases, and an optimal estimator that minimizes variance is developed. Variance estimation is also discussed for both cases. Numerous examples of surveys conducted at Statistics Canada illustrate the unification of both cases.

Lavallée and Caron investigate the problem of producing estimates when using record linkage methods to link two populations together. In particular, they consider the problem of producing estimates for one of the populations using a sample from the other one, assuming the two populations have been linked together. The Generalized Weight Share method is adapted to take into account the linkage weights in three different ways: (1) all links where the linkage weight is non-zero; (2) all links where the linkage weights are greater than a given threshold; and (3) the links are randomly chosen. These proposed estimators are compared with the classical approach through a simulation study.

Merkouris considers the problem of producing cross-sectional estimates with data collected from multiple panel surveys. Coverage of the cross-sectional population maybe incomplete due to individuals leaving or entering the population after the selection of the panel. By recognizing that a repeating panel survey is a special type of multiple frame survey, Merkouris is able to propose weighting strategies suitable for various multiple panel surveys. These weighting procedures can be used to combine information from the multiple panels to produce cross-sectional estimates that take into account the dynamic character of the multiple panel design.

Marker investigates survey design strategies to improve the quality of direct small area estimators, thus reducing the need for indirect, model-based estimators. Factors considered include stratification and oversampling, combining data from repeated surveys, harmonizing across different surveys, supplemental samples, and improved estimation procedures.

In their paper, Saigo, Shao and Sitter address the important problem of variance estimation under imputation for missing data. In their paper, they propose a bootstrap method that works for both smooth and non-smooth statistics, even for the case where the number of sampled clusters is small. This improves on their previously proposed bootstrap method which could suffer from serous overestimation when the number of sampled clusters is small. In addition to a bootstrap method, Saigo, Shao and Sitter also propose a repeated Balanced Repeated Replication method that captures the imputation variance in the presense of random imputation. These methods are illustrated through a simulation study.

Bellhouse and Stafford consider nonparametric local polynomial regression as an exploratory data analysis tool for data from complex surveys. They consider a single continuous regressor variable x, which is binned into a finite number of possible values, which may correspond to the precision of measurement of x, but may also be chosen otherwise. Point estimates of the local regression function, and associated variance estimates, are developed. The method is illustrated with an analysis of body mass indices from the Ontario Health Survey, and the nonparametric estimates are compared to those obtained from a parametric model.

In the final paper of this issue, Silva and Smith use a state space approach for modelling of compositional time series using data from a repeated complex survey. A compositional time series is a multivariate time series of proportions constrained to add to one at each time point. They first transform the data using an additive logistic transformation, and then model the transformed series. Estimation methods based on the Kalman filter are developed and then applied to data from the Brazilian Labour Force Survey. The Kalman filter also provides model-based estimates of variance and confidence limits for the transformed series. Estimates of trends and seasonal effects are compared to those obtained using X-11 ARIMA, and found to be generally smoother since they explicitly account for sampling errors in the raw estimates of the series.

M.P. Singh

Survey Methodology, December 2001
Vol. 27, No. 2, pp. 121-130
Statistics Canada

Using Propensity Scores to Control Coverage Bias in Telephone Surveys

KRISTIN BLENK DUNCAN and ELIZABETH A. STASNY[1]

ABSTRACT

Telephone surveys are a convenient and efficient method of data collection. Bias may be introduced into population estimates, however, by the exclusion of nontelephone households from these surveys. Data from the U.S. Federal Communications Commission (FCC) indicates that five and a half to six percent of American households are without phone service at any given time. The bias introduced can be significant since nontelephone households may differ from telephone households in ways that are not adequately handled by poststratification. Many households, called "transients", move in and out of the telephone population during the year, sometimes due to economic reasons or relocation. The transient telephone population may be representative of the nontelephone population in general since its members have recently been in the nontelephone population.

This paper develops a weighting adjustment for transients in an effort to reduce the bias due to noncoverage while controlling the increase in variance due to weighting. We use a logistic regression model to describe each household's propensity for transience, using data collected from a survey of distressed and non-distressed regions of Kentucky, Ohio, and West Virginia. Weight adjustments are based on the propensity scores. Estimates of the reduction in bias and the error of estimates are computed for a number of survey statistics of interest, using the propensity based weight adjustments and several alternative weight adjustments. The error in adjusted estimates is compared to the error of the standard estimate to assess the effectiveness of the adjustment.

KEY WORDS: RDD survey; Weight adjustments; Non-sampling error.

1. INTRODUCTION

The telephone is a standard mode of communication in today's world, and hence it is extremely useful for conducting surveys. Telephone surveys have come into use more and more as a growing percentage of people have phone connections. Most people who belong to the population that a survey seeks to make inferences about, the survey's target population, can be reached by phone. Therefore, the sample is drawn from the set of all people in households reachable through residential phone numbers. However, this sampling frame excludes all the people without telephone service who may compose a significant portion of some populations. It is currently estimated that in the United States, five and a half to six percent of households are without telephone service at any given time (Belinfante 2000). People without phone service tend to be different from people with service, particularly with regards to economic factors (Smith 1990). Results of the survey will not truly reflect the entire population if these differences are significant on matters of importance to the survey. The coverage bias is particularly troublesome in surveys that examine subgroups of the population with lower telephone penetration rates. These groups include people in lower income households and people who have not obtained a high school degree.

Poststratification on demographic variables associated with telephone coverage is helpful for reducing the coverage bias, but it does not completely solve the problem (Massey and Botman 1988). Another way to account for this coverage bias is to let people who are currently without telephone service be represented by people in the survey who have not had continuous service recently. People whose phone status has changed within the last year are referred to as transients. Transients move in and out of the telephone population, possibly for economic reasons, or service interruptions during relocation. Transients who currently have phone service may be good representatives of the nontelephone population because they are included in the sampling frame, yet they have recently been part of the nontelephone population.

A weighting adjustment suggested by Brick, Waksberg and Keeter (1996) uses transients in the sample to represent the nontelephone population. They use data from the U.S. Current Population Survey (CPS) to estimate unbiased weighting class adjustments for the transient respondents in their survey. Frankel, Ezzati-Rice, Wright and Srinath (1998) also employ this weighting class adjustment, and consider two similar adjustments. Brick, Flores Cervantes, Wang and Hankins (1999) and Frankel, Srinath, Battaglia, Hoaglin, Wright and Smith (1999) evaluate these adjustments using surveys that ask questions about telephone service, but that are not subject to telephone coverage bias. These studies found that employing weight adjustments based on transient status generally led to improved estimates.

This article studies an alternative method for computing a transient weight adjustment. Our method develops a model for predicting transience using demographic variables. The weight adjustment is then based on the

[1] Kristin Blenk Duncan and Elizabeth A. Stasny, Department of Statistics, Ohio State University, Columbus, OH 43210-1247.

respondent's propensity for transience. We also compare our propensity method to the method suggested by Brick *et al.* (1996), and to a response probability method where the weight adjustment is based on the length of interruption in telephone service.

We use data from the Appalachian Poll, an RDD telephone survey conducted by the Ohio State University's Center for Survey Research during June and July of 1999. The survey was sponsored by *The Columbus Dispatch*, and compared distressed and non-distressed regions of Kentucky, Ohio, and West Virginia. The study gathered information on quality of life issues and perceptions about the Appalachian regions, and also posed a series of standard demographic questions. A stratified sample was used, and just over 400 surveys were completed from each of the six strata (Appalachian and non-Appalachian regions of Ohio, Kentucky, and West Virginia). The poll targeted English speaking adults, 18 years of age or older, residing in the three states. Coverage bias is of particular concern in this survey since telephone coverage rates are lower than usual in the distressed Appalachian regions.

In section 2, we report on the literature describing telephone and transient populations. In this section we also explore differences between these groups in our data, illustrating the concern about coverage bias. Section 2 ends with our proposed model for predicting transience. Section 3 details the various weighting procedures. In section 4 we discuss the trade-off between bias reduction and increased variance from adjusted weights, and compare the weighting schemes. The final section summarizes the findings.

2. NONTELEPHONE AND TRANSIENT TELEPHONE POPULATIONS

The target population for a telephone survey can be categorized by telephone status into four groups: continuous service households, transient households which are currently with service, transient households which are currently without service, and chronic nontelephone households. We need to know something about the size of each of these groups in order to account for coverage bias in the survey. Data from the FCC is useful for examining long term trends in the size of the nontelephone population. Not as much is known, however, about the short-term changes in phone coverage.

Keeter (1995) used panel surveys to study the dynamics of the transient phone population. In the March 1992 and 1993 CPS, it was found that 94.1% of households in the sample at both times had a phone at both time points, 2.6% at neither point, and 3.4% had a phone at one interview, but not the other. Fifty-seven percent of respondents who reported having no phone at either interview were transient. If the measurements could be taken continuously, rather than at two points in time, even more households would be labeled transient. Keeter concludes that, "a sizable minority of nontelephone households, at the least, have recently been in the telephone population or are soon to join it. Such transient households constitute a measurable segment of telephone households and thus can provide data to characterize the nontelephone population," (Keeter 1995, page 201). The same article asserts that, "Transient telephone households are much more like nonphone households than those with continuous service," (Keeter 1995, page 209). This conclusion is based on formal tests using demographic variables from the CPS. Data from the National Survey of America's Families presented in Brick *et al.* (1999) supports Keeter's findings. Since transients make up a nontrivial proportion of the nontelephone population and transients are more similar to the nontelephone households than they are to continuous service households, it is reasonable to use data from the transients in the sample to attempt to reduce coverage bias.

In the Appalachian Poll, 140 of the 2,463 respondents, or 5.7%, replied positively to the question, "During the last twelve months has your household ever been without telephone service for one week or more?" These respondents are categorized as transients. In the Appalachian regions, the transience rate is 7.4% while the rate is only 3.9% in non-Appalachian regions.

Table 1 compares transient and nontransient households from the sample in regards to selected variables. The large differences between the two populations illustrate the need for bias reduction. People who live in transient households are much younger, have lower incomes, and they are less likely to be employed full time. They also have less access to health insurance and computers.

Table 1
Selected Characteristics of Nontransient and Transient Households

Characteristics	Nonstransient	Transient
Median Age	47.0	37.5
Household income Less than $20K	27.8%	60.0%
Employed full-time or retired	55.0%	34.5%
No health insurance	12.7%	30.0%
Owns or is buying residence	79.4%	61.4%
Computer in home	47.4%	26.4%
Not enough money for food	12.3%	42.9%

Note: Statistics are based on unweighted frequencies in the sample which oversampled from the Appalachian regions, and thus are not representative of population quantities.

A model for transience. Using the Appalachian Poll sample, we develop a logistic regression model to predict transience with demographic variables. The independent variables used to predict transience are age, employment status, race, income, and region. The model is described in the Appendix. Education and tenure are also good predictors of transience, but they are strongly correlated with

the other variables in our model, and thus, we chose not to include them. For a comparison of models that predict telephone coverage, see Smith (1990). We will use our model in the propensity weighting adjustment described in the following section.

3. WEIGHT ADJUSTMENTS

We consider several weighting schemes that attempt to account for the coverage bias inherent in telephone surveys. Each of these schemes is compared to the actual weighting procedure used for the Appalachian Poll. In the standard procedure, a base weight was calculated for each respondent. This adjustment is (# adults in household) / (# voice telephone lines), or the inverse of the respondent's probability of being in the sample. Then weights were raked in each of the six strata to agree with 1990 Census proportions for age group, education level, and gender. Finally, the weights were scaled to the sample sizes within the six strata.

3.1 Length of Disconnect

Respondents to the Appalachian Poll who replied "yes" to the question about an interruption in phone service of one week or longer were then asked how many days they were without service in the last year. A simple approach to the coverage bias problem is to give transients a weight adjustment inversely proportional to the fraction of the year that they were with service. For example, a person who has only had service for six months out of the last twelve receives a weight of two, thus representing himself and one other person in the population with a six-month disconnect who is currently without service.

This naïve approach is included in the analysis for comparison with other schemes. It is referred to as the day scheme (DAY). Weight adjustments are calculated as 365/(365 – # days without service). This weight adjustment is applied after the base weight described above, and before the weights are raked.

While this approach is logical, it is not practical for controlling variance. It is usually considered undesirable to use weighting factors larger than three. In fact, for many large surveys conducted by the U.S. Census Bureau, if weighting factors are larger than two, respondents are merged into larger groups and a group weight is calculated in order to obtain lower weighting-adjustment factors; see, for example, CPS (1978).

This simple approach becomes more practical when respondents are grouped by the length of their interruption in service. In a scheme called day group (DAYG), transients are grouped into quartiles across the entire sample by length of interruption in phone service. These quartiles correspond to interruptions of one week, more than one week but less than three weeks, three weeks to two months, and more than two months. The weight adjustment for each group is 365/(365 – avg. # days without service), and it is also applied after the base weight, prior to raking. This grouping procedure is helpful for reducing the variance caused by extremely long interruptions.

3.2 Weighting Class Adjustment Scheme

Brick *et al.* (1996) also implement a response probability adjustment to reduce coverage bias. Under their procedure, they partition the target population into the four components described in section 2: t_1 is the number of persons living in continuous service households; t_2 is the number of persons living in transient households that currently have service; t_3 is the number of persons living in nontelephone households that have not had any service in the last year; and t_4 is the number of persons living in transient households that are currently without service. The response probability model the authors use assumes that $t_3 = 0$. With this assumption, an unbiased weight adjustment is $A = (t_2 + t_4)/t_2 = 1 + (t_4/t_2)$, the inverse of the proportion of the transient population that currently has service. Unfortunately, these population quantities are unknown and must be estimated. Following the lead of Brick *et al.*, we use CPS data to estimate $t_1 + t_2$, the number of persons who currently have service, and t_4; call these estimates $\hat{t}_1 + \hat{t}_2$ and $\hat{t}_4$, respectively. From the Appalachian Poll, separate estimates of t_1 and t_2 are available; designate these estimates as t_1^* and t_2^*, respectively. Since the estimates come from different surveys, ratios are used in the weight adjustment, and A is estimated by

$$A' = 1 + \frac{\dfrac{\hat{t}_4}{\hat{t}_1 + \hat{t}_2}}{\dfrac{t_2^*}{t_1^* + t_2^*}}. \qquad (1)$$

Some persons are more likely to live in nontelephone households than others, so Brick *et al.* classified transients into cells based on characteristics associated with not having a telephone, and computed the weight adjustment for each cell. Four classification schemes, which categorized respondents by either education or tenure, length of interruption, and race/ethnicity were considered.

Brick *et al.* found schemes that classified respondents as transients if they had an interruption of one week or more to be superior to schemes that used a cut-off of one month, so for the Appalachian Poll data we use the one-week cut-off. Due to the small number of Hispanics in the Appalachian Poll sample, we do not categorize by ethnicity. Thus, for our analyses, the cell classifications for two schemes that use the method described by Brick *et al.* (1996) are defined as follows:

> **BWKE** – households that had a service interruption of one week or more within categories defined by education (less than high school, high school diploma, college diploma or above) and race (black, non-black); and

BWKT – households that had a service interruption of one week or more within categories defined by tenure (own/other, rent) and race.

The disadvantage of using these schemes in our study is that the estimates needed from the CPS are available by state, but not by region since the CPS does not sample from all counties. Persons in Appalachian regions are less likely to have telephones, but we cannot account for this with the available CPS data. Even when we consider statewide data, the sample size of the CPS is not large enough to get reliable values of $\hat{t}_4$ in all of the cells. For example, in 1999 the CPS did not sample any blacks with a college degree or higher who live in Kentucky and do not have telephone service. Thus, the weighting cell adjustments computed for use with the Appalachian Poll are based on CPS data from the three states combined.

3.3 Raking Ratio Adjustment

Lohr (1999) explains the use of raking ratio estimates to adjust for nonresponse in surveys. We propose a similar use of raking to account for coverage bias. We estimate the proportion of the population with continuous telephone service, and then use raking to allow transients in the sample to represent the portion of the population without continuous telephone service.

The percent of households without continuous service is estimated by

$$1 - \left(\frac{\bar{t}_1 + \bar{t}_2}{\tilde{t}_1 + \tilde{t}_2 + \tilde{t}_4}\right)\left(\frac{t_1^*}{t_1^* + t_2^*}\right), \qquad (2)$$

where $\bar{t}_i, i = 1, 2, 4$, is obtained from the FCC data. The first fraction estimates the proportion of households that currently have service, and the second fraction estimates the proportion of nontransient households among households with service. Again, we assume that $t_3 = 0$. The FCC gives telephone penetration rates by state, but not by region. Data from the 1990 Census does give penetration rates by county, but rates changed from 1990 to 1999. Therefore, to estimate the 1999 regional penetration rate, we maintained a constant ratio of percent of households without a phone in the non-Appalachian regions to percent of households without a phone in the Appalachian regions and adjusted the 1990 Census regional rates to match the 1999 state rates. Table 2 gives the data we used to compute the 1999 state rates, and the resulting estimates.

In a scheme referred to as transient raking, or TRAK, transient status is included as a control variable for raking along with age, gender, and education level. The totals we used for raking by transient status are given in Table 2.

3.4 A New Propensity Weighting

An estimated propensity score is sometimes used to create a weight adjustment to account for nonresponse in surveys where some variables are known for the nonrespondents. For example, in a face-to-face household interview the interviewer knows the address of the nonrespondent and may have information about the person's race, gender, and age. A logistic regression model that describes propensity for response is developed, and respondents are assigned a weight of $1/\hat{p}$, where $\hat{p}$ is the estimated propensity to respond (Little and Rubin 1987). This procedure gives higher weights to sampled households that are more similar to the nonrespondents. Since there is typically no data on the excluded nontelephone population in telephone surveys, a modified approach is taken to using a propensity score. We only adjust the weights for the transients since they will represent the missing part of the sample: weights for nontransients remain unadjusted. The

Table 2
Computation of Transient Status Raking Totals

	Kentucky		Ohio		West Virginia	
	Ap	Non-Ap	Ap	Non-Ap	Ap	Non-Ap
Appalachian Poll Data						
Sample Size	412	407	413	405	411	415
# transients in sample	38	19	18	13	36	16
Percent of sample without cont. service	9.2	4.7	4.4	3.2	8.8	3.9
Census and FCC Data						
1990 State % no phone	10.2	10.2	4.7	4.7	10.3	10.3
1990 Region % no phone	19.1	8.2	11.7	4.5	14.3	8.4
1999 State % no phone	6.7	6.7	5.2	5.2	7.3	7.3
Percent of state pop. living in region	18.6	81.4	2.6	97.4	31.8	68.2
Estimates						
Ratio of Non-Ap to Ap noncoverage	0.429	0.429	0.385	0.385	0.587	0.587
Estimated 1999 region % no phone	12.5	5.4	13.0	5.0	10.1	6.0
Estimated % of pop. without cont. service	20.6	9.8	16.7	8.1	18.0	9.6
Desired # of transients in sample	85	40	69	33	74	40

weight adjustment for transients is $1/(1-\hat{p})$, where $\hat{p}$, the estimated propensity for transience, is described by the model in section 2.1. Households with a higher estimated propensity for transience may be more representative of the nontelephone population and they receive higher weight adjustments. This adjustment is applied to the base weight, and the scheme is called propensity (PROP).

Transience is not that common, and most estimated propensity scores are fairly low. In the PROP scheme, the average weight adjustment for a transient household is 1.167. This adjustment is not large enough for transients to represent themselves and the entire nontelephone population. That is, when the weights are scaled to sum to the population size, the sum of the final weights for transients is less than the size of the transient population. To account for this under-representation, the propensity weight adjustment is applied, and then transient is used as a control variable for raking along with age, education, and gender. The estimated population sizes for transients are computed as in section 3.3. This weighting scheme is called augmented propensity, or AUGP.

4. FINDINGS

The analysis and comparison of the adjustment schemes presented here parallels the analysis performed by Brick *et al.* (1996). We first discuss the change in variance resulting from adjusting the weights to reduce coverage bias and present a statistic for measuring the relative variability. Then, the schemes are evaluated by comparing the variance of adjusted estimates to the mean squared error of the standard estimate.

4.1 Changes in Variability

The goal of the adjustment schemes is to decrease coverage bias while controlling variance. Adjustment of the weights to reduce the bias increases the variability of the weights, hence increasing the variance of the estimates. Kish (1992) gives a formula for measuring the increase in variance due to unequal weights. Brick *et al.* (1996) refer to this expression as the variance inflation factor (VIF). The VIF can be written as

$$\text{VIF} = 1 + [\text{CV}(\text{weights})]^2, \qquad (3)$$

where CV(weights) is the coefficient of variation of the weights. A VIF ratio is computed to compare the VIF of a new weighting scheme to that of the standard weighting scheme. Table 3 gives VIF ratios for the six strata in the Appalachian Poll data under each scheme described in section 3. A VIF ratio of 1.12, for example, indicates an increase in variance of 12 percent over the variance using the standard weighting scheme. The VIF ratio values are reasonable for all schemes except the DAY scheme which sees an average variance increase of 300 percent. The VIF ratio values for our PROP scheme are all very close to one, suggesting that the PROP weight adjustments will not increase the variance of our estimates.

4.2 Coverage Bias Reduction

Estimates of seventeen population proportions using survey variables from the Appalachian Poll were calculated for the standard weighting procedure and for each of the seven adjustment schemes (see Table 4 for a list of the seventeen variables). WesVar software was used to calculate standard errors for these estimates by means of replication. We would like to assess the effectiveness of each scheme for reducing the coverage bias on these seventeen characteristics. Estimates from an independent source that are free of telephone coverage bias would be ideal for such an assessment. Unfortunately, such benchmarks are unavailable and some model assumptions are necessary in order to perform an evaluation. We assume that the weight adjustment procedures reduce the coverage bias. Thus the difference between the standard estimate and the adjusted estimate is considered to be an unbiased estimate of the decrease in coverage bias resulting from the adjustment. The assumption favors the adjusted estimates, considering them to be unbiased.

Table 3
Ratios of Variance Inflation Factor Due to Weight Adjustment

Region	Ratio of scheme's VIF to standard weight's VIF						
	DAY	DAYG	BWKE	BWKT	TRAK	PROP	AUGP
Non-Appalachian Ohio	0.999	0.997	1.004	1.023	1.063	0.999	1.061
Appalachian Ohio	1.480	1.016	1.039	1.091	1.331	0.999	1.336
Non-Appalachian Kentucky	4.151	1.040	1.018	1.054	1.030	0.999	1.029
Appalachian Kentucky	2.433	1.069	1.045	1.042	1.129	1.003	1.145
Non-Appalachian West Virginia	6.331	1.027	1.010	1.029	1.020	0.999	1.024
Appalachian West Virginia	2.935	1.085	1.058	1.053	1.116	1.005	1.119
Scheme Average	3.055	1.039	1.029	1.049	1.115	1.001	1.119

Table 4
Estimated Reduction in Bias and Bias Ratio for Selected Characteristics

Characteristic	Standard estimate		Estimated reduction in bias							Bias Ratio						
	Estimate	St. error	DAY	DAYG	BWKE	BWKT	TRAK	PROP	AUGP	DAY	DAYG	BWKE	BWKT	TRAK	PROP	AUGP
Owns Home																
Non-Appalachian Ohio	72.2	3.1	0.6	0.5	0.5	1.2	1.4	0.1	1.6	0.2	0.2	0.2	0.4	0.5	0.0	0.5
Appalachian Ohio	75.4	2.8	4.4	0.6	0.6	2.1	3.2	0.3	3.5	1.6	0.2	0.2	0.8	1.1	0.1	1.2
Non-Appalachian Kentucky	68.6	3.1	7.2	0.8	0.9	1.8	1.5	0.2	1.5	2.3	0.3	0.3	0.6	0.5	0.1	0.5
Appalachian Kentucky	80.5	2.2	2.9	0.8	0.3	1.3	0.3	0.0	0.3	1.3	0.3	0.1	0.6	0.1	0.0	0.1
Non-Appalachian West Virginia	80.0	2.3	14.2	1.6	0.9	1.9	1.4	0.2	1.4	6.1	0.7	0.4	0.8	0.6	0.1	0.6
Appalachian West Virginia	81.9	2.2	8.2	0.7	-0.4	0.5	-0.3	0.0	-0.2	3.7	0.3	-0.2	0.2	-0.1	0.0	-0.1
No Health Insurance																
Non-Appalachian Ohio	7.3	1.7	0.0	-0.1	-0.6	-1.4	-1.7	-0.1	-1.8	0.0	-0.1	-0.4	-0.8	-1.0	-0.1	-1.1
Appalachian Ohio	12.6	2.1	0.9	0.1	0.3	0.3	0.5	0.1	0.6	0.4	0.1	0.1	0.2	0.3	0.0	0.3
Non-Appalachian Kentucky	8.8	1.8	1.8	0.4	0.2	0.3	0.0	0.1	0.1	1.0	0.2	0.1	0.2	0.0	0.0	0.0
Appalachian Kentucky	22.2	2.4	3.4	0.1	-0.1	-0.2	-0.8	-0.4	-1.5	1.4	0.0	0.0	-0.1	-0.3	-0.2	-0.6
Non-Appalachian West Virginia	14.2	2.1	-4.8	-0.5	-0.7	-1.0	-1.2	-0.3	-1.4	-2.3	-0.2	-0.3	-0.5	-0.6	-0.1	-0.7
Appalachian West Virginia	24.6	2.5	2.5	-0.8	-1.7	-1.3	-2.7	-0.6	-3.0	1.0	-0.3	-0.7	-0.5	-1.1	-0.2	-1.2
Not enough Money for Food																
Non-Appalachian Ohio	10.8	1.9	-0.7	-0.6	-0.9	-1.6	-2.2	-0.1	-2.1	-0.4	-0.3	-0.5	-0.9	-1.2	0.0	-1.2
Appalachian Ohio	16.2	2.5	-4.7	-0.8	-0.6	-1.3	-3.3	-0.2	-3.4	-1.9	-0.3	-0.3	-0.5	-1.3	-0.1	-1.4
Non-Appalachian Kentucky	11.4	2.4	-3.3	-0.8	-1.3	-1.7	-1.6	-0.4	-1.8	-1.4	-0.3	-0.5	-0.7	-0.7	-0.2	-0.8
Appalachian Kentucky	20.2	2.4	-7.4	-2.3	-2.1	-2.1	-3.8	-0.4	-3.8	-3.1	-1.0	-0.9	-0.9	-1.6	-0.2	-1.6
Non-Appalachian West Virginia	14.0	2.1	4.3	-0.1	-1.0	-1.4	-1.7	-0.3	-1.8	2.1	0.0	-0.5	-0.7	-0.8	-0.2	-0.9
Appalachian West Virginia	16.4	2.0	1.5	-0.7	-1.0	-0.9	-2.2	-0.5	-2.6	0.8	-0.3	-0.5	-0.4	-1.1	-0.3	-1.3
Computer in Home																
Non-Appalachian Ohio	60.1	3.0	0.4	0.3	0.6	1.2	1.3	0.1	1.4	0.1	0.1	0.2	0.4	0.5	0.0	0.5
Appalachian Ohio	40.0	3.0	1.2	0.2	0.3	0.8	1.8	0.1	2.0	0.4	0.1	0.1	0.3	0.6	0.0	0.7
Non-Appalachian Kentucky	44.5	3.0	6.7	0.9	0.8	1.1	0.9	0.2	1.0	2.3	0.3	0.3	0.4	0.3	0.1	0.3
Appalachian Kentucky	29.7	2.3	1.9	1.0	0.9	1.1	2.3	0.0	1.9	0.8	0.4	0.4	0.5	1.0	0.0	0.8
Non-Appalachian West Virginia	46.2	2.6	7.6	0.6	1.1	1.2	1.5	0.3	1.6	2.9	0.2	0.4	0.4	0.6	0.1	0.6
Appalachian West Virginia	36.1	2.7	4.3	1.0	0.3	0.4	0.2	0.3	0.5	1.6	0.4	0.1	0.2	0.1	0.1	0.2
Summary of Seventeen Variables																
Mean absolute value			0.032	0.005	0.006	0.009	0.013	0.002	0.014	1.396	0.235	0.620	0.412	0.885	0.075	0.885
Median absolute value			0.022	0.005	0.006	0.011	0.014	0.001	0.014	0.995	0.240	0.245	0.420	0.605	0.055	0.665

Note: In addition to the four proportions listed in the table, the summary of seventeen variables includes worry about income, better off economically in the 1990's, dissatisfied with own net worth, married, have children, unemployed, college graduate, in good or excellent health, serious illness in household, no family doctor, satisfied with own housing, very safe drinking water, and internet access in home.

Using our assumption, we compare the estimate from each scheme to the standard estimate. The reduction in coverage bias is estimated by the difference between the standard estimate and the adjusted estimate. There are seven different estimates of the bias reduction, one for each scheme. The estimated reduction in bias is given by

$$b_i = \hat{p}_s - \hat{p}_i, \tag{4}$$

where b_i is the estimated bias reduction using scheme i, $\hat{p}_s$ is the standard estimate, and $\hat{p}_i$ is the estimate from adjustment scheme i. Estimated reductions in bias for four characteristics by the six strata are given in Table 4 for each scheme. For the characteristics owns home, not enough money for food, and computer in home, the direction of the bias is fairly consistent across schemes and regions. Reassuringly, the bias is in the expected direction for these characteristics, with fewer people owning homes, more people not having enough money for food, and fewer people having computers in their homes, than is indicated by the estimates using the standard weighting scheme. For health insurance, the direction of the bias is mostly consistent across regions. The standard estimate is biased upward for Appalachian Ohio and non-Appalachian

Kentucky, and generally biased downward in the other regions.

The absolute size of the reduction in bias by itself is not fully meaningful, because it does not account for the amount of sampling error associated with the estimate. Therefore, we also calculate the bias ratio, as in Brick *et al.* (1996). The bias ratio for scheme i, r_i, is given by

$$r_i = \frac{b_i}{\text{se}(\hat{p}_s)}, \quad (5)$$

where $\text{se}(\hat{p}_s)$ is the standard error of the standard estimate. Table 4 also gives the bias ratio for the selected estimates. DAY, TRAK, and AUGP give the largest bias ratios; for these adjustment schemes the bias is not negligible when we consider the standard error. DAYG and PROP have low bias ratios, indicating that the bias reduction is small compared to the error of the estimate.

4.3 Mean Square Error

Since the standard estimates are thought to be biased, error should be measured with mean square error rather than variance. The MSE of the standard estimate is approximated by

$$\text{mse}_i = \text{var}(\hat{p}_s) + b_i^2 \quad (6)$$

for each adjustment scheme. Recall that we are assuming the adjusted estimates are unbiased, so that the mean square errors of these estimates are equal to their variances. The variance of the adjusted estimates can be approximated by two methods. The first approximation is obtained by multiplying the VIF ratio in Table 3 by the variance of the standard estimate. Alternatively, we can use the variance of the adjusted estimate obtained from replication methods.

The error of the adjusted estimate is compared to the error of the standard estimate in the mean square ratio (MSR). Using the VIF variance, the estimated MSR is given by

$$\text{msr}_{\text{VIF}i}(\hat{p}) = \frac{100 \times \text{VIF Ratio}_i \times \text{var}(\hat{p}_s)}{\text{mse}_i(\hat{p})}. \quad (7a)$$

For the replication variance, the estimated MSR is given by

$$\text{msr}_{\text{VAR}i}(\hat{p}) = \frac{100 \times \text{var}_i(\hat{p})}{\text{mse}_i(\hat{p})}, \quad (7b)$$

where $\text{var}_i(\hat{p})$ is the estimated variance of the adjusted estimate, obtained through replication. An MSR of 100 indicates that the variance of the adjusted estimate is exactly equal to the mean squared error of the standard estimate. An MSR above 100 means the variance of the adjusted estimate is larger than the MSE of the standard estimate, and the bias/variance trade-off for the scheme is not favorable. An MSR below 100 means that the adjusted estimate is an improvement over the standard estimate in terms of overall error.

Table 5 gives estimated MSR values for selected survey variables from the Appalachian Poll, and a summary of these values for seventeen variables from each adjustment scheme. The MSR estimates vary between regions and between schemes. The msr values computed using the two different variances also differ, but the summary values are similar for both variances. The DAY scheme has the highest msr values, indicating that this weight adjustment is not worthwhile because it increases the variance too much. TRAK and AUGP have the lowest mean and median msr values, though these schemes produced unfavorable estimates for a few characteristics as indicated by the high maximum msr values. The weighting class adjustment schemes BWKE and BWKT performed well and their maximum estimated mean square ratio values are fairly low. All of the msr values for the PROP scheme are near 100, suggesting that the overall error in estimates computed with this scheme is comparable to the error in the standard estimates.

5. CONCLUSIONS

While telephone use is commonplace, telephone surveys will always contain some bias since nontelephone households are excluded from the sampling frame, and the nontelephone population has characteristics that differ from those of the telephone population. Coverage bias is alleviated by poststratification on variables such as income and education and may not be a problem in some instances. However, for surveys that target poor or rural areas where telephone penetration rates are lower, the coverage bias is a large concern.

We have proposed a few new methods for reducing the coverage bias by adjusting the weights of respondents in the transient population. We compared the resulting estimates to those from other existing methods. In the analysis of these methods, it was assumed that the adjusted estimates are unbiased. In the absence of unbiased benchmark estimates this assumption cannot be validated. The mean square ratios presented here are likely to be biased downward since the bias of the adjusted estimate is not included. The estimated MSR is still useful for comparing methods, however, and gives a good measure of the effectiveness of the weight adjustments.

As anticipated, the DAY method was found to have too much variability to be useful. The day group (DAYG) method appears to perform better, but most of the mean square ratios for this scheme are close to 100, meaning that we do not see a large improvement over the standard estimate. The advantage of this scheme lies in its simplicity. The weight adjustment is easy to apply and does not require auxiliary data.

Table 5
Mean Square Ratio for Selected Characteristics

Characteristic	VIF Mean Square Ratio							Variance Mean Square Ratio						
	DAY	DAYG	BWKE	BWKT	TRAK	PROP	AUGP	DAY	DAYG	BWKE	BWKT	TRAK	PROP	AUGP
Owns Home														
Non-Appalachian Ohio	96.1	97.2	97.3	88.1	87.5	99.8	84.5	98.6	98.2	98.2	88.4	81.8	99.9	78.7
Appalachian Ohio	42.3	97.4	98.7	68.9	57.9	99.1	52.5	71.7	96.5	89.6	71.1	51.6	99.2	46.6
Non-Appalachian Kentucky	63.9	97.6	94.5	77.7	83.1	99.4	83.5	21.9	98.2	92.8	75.7	81.3	98.8	80.1
Appalachian Kentucky	89.3	96.1	102.3	77.0	110.6	100.3	112.4	116.0	100.7	104.1	88.4	119.2	100.0	118.3
Non-Appalachian West Virginia	16.6	71.2	89.1	62.3	75.8	99.2	75.5	28.6	81.1	94.4	71.4	85.5	100.1	84.2
Appalachian West Virginia	20.2	98.4	103.2	100.3	109.3	100.5	110.6	43.5	106.0	101.1	104.8	108.8	99.1	108.9
No Health Insurance														
Non-Appalachian Ohio	99.9	99.0	88.2	61.4	51.9	99.5	48.5	98.8	100.5	112.1	101.7	82.0	101.9	76.1
Appalachian Ohio	126.8	101.3	102.1	106.5	125.1	99.9	123.5	92.3	98.8	95.6	94.0	105.8	99.0	100.4
Non-Appalachian Kentucky	206.4	99.9	100.9	102.8	103.0	99.8	102.7	39.0	87.9	90.7	86.2	97.8	96.6	95.5
Appalachian Kentucky	82.7	106.7	104.4	103.7	102.1	97.9	84.3	53.5	109.9	104.9	105.3	114.1	100.0	100.5
Non-Appalachian West Virginia	100.2	97.1	90.6	84.0	77.7	97.9	71.3	136.6	99.2	94.0	89.7	90.6	100.8	84.3
Appalachian West Virginia	149.6	99.2	74.1	83.5	52.8	95.7	46.7	107.0	96.5	75.1	84.3	51.5	96.7	45.4
Not enough Money for Food														
Non-Appalachian Ohio	86.5	90.5	80.5	57.9	45.2	99.7	45.6	105.2	100.8	104.3	94.5	66.3	102.0	67.0
Appalachian Ohio	31.9	92.9	97.4	86.1	48.6	99.4	46.4	68.5	98.2	96.8	90.2	69.4	101.1	66.4
Non-Appalachian Kentucky	139.1	94.1	78.2	69.5	69.5	96.7	64.3	320.7	96.8	91.9	85.7	77.6	100.1	68.5
Appalachian Kentucky	22.3	55.8	57.6	58.7	31.0	97.2	31.9	30.5	68.4	68.5	69.5	36.8	100.2	38.5
Non-Appalachian West Virginia	117.3	102.6	82.4	71.0	59.6	97.7	57.2	105.7	101.9	94.7	88.3	71.5	101.6	68.5
Appalachian West Virginia	181.6	97.0	84.1	88.5	50.4	94.1	39.9	92.2	98.8	89.6	92.9	59.0	97.5	48.3
Computer in Home														
Non-Appalachian Ohio	98.1	98.5	96.4	88.2	88.1	99.8	86.1	99.5	99.5	102.0	102.1	106.3	100.6	102.8
Appalachian Ohio	127.2	101.2	103.1	101.2	96.2	99.7	92.5	116.0	99.6	101.2	96.5	94.1	99.1	86.5
Non-Appalachian Kentucky	67.7	94.9	94.4	92.7	93.6	99.5	92.8	27.1	93.7	91.5	89.7	90.3	98.2	88.4
Appalachian Kentucky	147.1	89.0	91.7	85.1	55.7	100.3	68.4	58.9	81.1	85.5	79.5	46.8	100.9	66.6
Non-Appalachian West Virginia	66.8	96.9	86.6	85.8	76.1	98.5	73.5	59.6	95.8	85.1	85.3	72.9	98.6	68.6
Appalachian West Virginia	82.7	95.6	104.4	103.0	111.2	99.6	108.2	41.8	88.1	101.6	99.9	113.3	98.3	107.0
Summary of Seventeen Variables														
Mean	137.6	97.5	94.3	92.2	85.2	99.3	83.8	125.2	99.1	97.1	96.8	96.0	100.2	93.5
Median	107.5	99.0	99.1	97.1	89.8	99.8	86.3	94.8	98.9	98.7	98.5	98.0	100.0	92.4
Minimum	10.9	55.8	0.9	57.9	4.1	94.1	5.7	7.0	68.4	43.1	62.1	7.6	94.6	6.0
Maximum	607.7	108.5	104.8	109.1	133.1	100.5	133.5	695.2	140.8	144.5	147.5	593.8	116.7	545.4
Percent below 100	47.1	60.8	61.8	58.8	65.7	87.3	67.6	63.7	62.7	56.9	58.8	53.9	58.8	58.8

Note: In addition to the four proportions listed in the table, the summary of seventeen variables includes worry about income, better off economically in the 1990's, dissatisfied with own net worth, married, have children, unemployed, college graduate, in good or excellent health, serious illness in household, no family doctor, satisfied with own housing, very safe drinking water, and internet access in home.

The weighting class adjustment schemes have the benefit of giving more weight to respondents in cells where the likelihood of having a phone is lower. For these schemes, greater bias reduction was seen in variables correlated with the classification variables. For example, home ownership and computer ownership are positively correlated, and the BWKT scheme, which classified respondents by home ownership, produced estimates of the percent of households with a home computer that were consistently lower than the standard estimates. Table 5 shows that the BWKE and BWKT schemes produce an improved estimate most of the time. It should also be noted that when these schemes produce an estimate that it not an improvement, the increase in variance remains fairly small. The weighting class adjustment method works well for samples of large populations, such as states or countries, since the outside data needed to compute the adjustments is readily available. The method is more difficult to use for very specific samples such as counties.

The raking ratio adjustment, TRAK, produced a number of very favorable estimated MSR values. With this scheme we were able to account for the difference in telephone penetration rates by region, but not the differences across other demographic characteristics. Variability was introduced when we estimated the regional rates from the state rates, thus, as with the weighting class adjustment, the scheme works better for samples of larger populations. While the mean and median estimated MSR values were low for this scheme, the scheme also produced some high mean square ratios. The higher ratios occurred in Ohio where the percent of transients in the sample was low compared to the estimated percent without continuous service.

The propensity adjustment alone, PROP, provided too little reduction in bias to be worthwhile. The propensity adjustment is advantageous, however, because it allows us to account for differences in the likelihood of having telephone service without using outside data. When used in conjunction with raking, the propensity based scheme AUGP produced good results.

There are many issues to consider when determining which adjustment scheme is preferred. As mentioned previously, the weighting class adjustment schemes BWKE and BWKT are difficult to implement if you have a very specific target population. These schemes are fairly conservative, however, in that they typically reduce the bias without increasing the variance. The schemes that employed raking usually performed better than the weighting class adjustment schemes, but the larger weight adjustments sometimes led to increased variances. It may be advisable to compute estimates using several schemes and then determine which scheme offers the best bias-variance trade-off.

Brick *et al.* (1996) note that these weight adjustments for telephone coverage should be more beneficial in reducing mean squared error when the sample size of the survey is large. As the sample size increases, the bias ratio increases since the bias is unaffected but the standard error of the estimate, which is in the denominator, decreases.

The findings suggested by this study and others indicate that the adjustments could be useful for many estimates from telephone surveys and should be seriously considered. The benefits of adjustment appear to outweigh the penalties in the weighting class adjustment schemes, the raking scheme, and the augmented propensity scheme. In light of the smaller sample size and special target population of the Appalachian Poll, generalizations of these findings should not be made until the methods receive further evaluation. These weight adjustments still need to be tested using a survey that is free of coverage bias, one that includes nontelephone households in the sampling frame and collects information on telephone status, in order to assess the validity of the assumptions. Data from the National Survey of America's Families, or the National Health Interview Survey may be appropriate for evaluating the adjustment methods and the assumptions.

ACKNOWLEDGEMENTS

This work was supported in part by a fellowship from the Center for Survey Research at the Ohio State University. We thank Dr. Paul Lavrakas and the Center for Survey Research for allowing us to use the Appalachian Poll data. We would also like to thank the referees for their helpful comments.

APPENDIX

Logistic Regression of Transient Status

Below is our model for predicting transient status. Most of the variables in the model relate to socioeconomic status. The coefficients indicate that young people, those with low income, those who are not employed full-time, American Indians and African Americans, and residents of distressed counties have higher propensities for transience. The high significance level of the Hosmer and Lemeshow test indicates a very good fit of the model. The large area under the ROC curve tells us that the model discriminates well.

Variable Coding

Age
- 0 - "Refused" (Count = 9)
- 1 - 18 to 29 years
- 2 - 30 to 44 years
- 3 - 45 to 59 years
- 4 - over 60

Low Income
- 0 - Household income over $20,000 or refused
- 1 - Household income under $20,000

Employment Status
- 0 - Employed full-time or retired
- 1 - Other (refused, part-time, housekeeper, student, unemployed, other)

Race
- 0 - Caucasian, Alaskan Native, Hispanic, or Asian
- 1- American Indian, African-American, Black, or other

Appalachian
- 0 - Does not live in a distressed county of KY, OH, or WV
- 1 - Lives in a distressed county

Kentucky/West Virginia
- 0 - Ohio
- 1 - Kentucky or West Virginia

Results

Variables in the Equation

Variable	B	S.E.
Age (Refused)	-2.107	12.160
Age (18-29)	2.006	0.357
Age (30-44)	1.664	0.347
Age (45-59)	1.064	0.364
Low Income	1.358	0.189
Employment Status	0.397	0.187
Race	1.136	0.292
Appalachian	0.531	0.196
KY/WV	0.567	0.216
Constant	-5.712	0.401

Hosmer and Lemeshow Goodness of Fit Test

Chi-Square	3.568
Degrees of Freedom	8
p-value	0.894

ROC Curve

Area under the Curve	0.782

REFERENCES

BELINFANTE, A. (2000). Telephone Subscribership in the United States. Industry Analysis Division, Common Carrier Bureau, Federal Communications Commisssion, Washington, D.C. 20554.

BRICK, J.M., FLORES CERVANTES, I., WANG, K. and HANKINS, T. (1999). Evaluation of the use of data on interruptions in telephone service. *Proceedings of the American Statistical Association Section on Survey Research Methods*, 376-381.

BRICK, J.M., WAKSBERG, J. and KEETER, S. (1996). Using data on interruptions in telephone service as coverage adjustments. *Survey Methodology*, 22, 185-197.

CURRENT POPULATION SURVEY (1978). Current Population Survey: Design and Methodology. Technical Paper 40. Department of Commerce, Bureau of the Census, Washington, D.C.

FRANKEL, M.R., EZZATI-RICE, T., WRIGHT, R.A. and SRINATH, K.P. (1998). Use of data in interruptions in telephone service for noncoverage adjustment. *Proceedings of the American Statistical Association Section on Survey Research Methods*, 290-295.

FRANKEL, M.R., SRINATH, K.P., BATTAGLIA, M.P., HOAGLIN, D.C., WRIGHT, R.A. and SMITH, P.J. (1999). Reducing nontelephone bias in RDD surveys. *Proceedings of the American Statistical Association Section on Survey Research Methods*, 934-939.

KEETER, S. (1995). Estimating noncoverage bias from a phone survey. *Public Opinion Quarterly*, 59, 196-217.

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

LITTLE, R., and RUBIN, D. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley and Sons, 55-60.

LOHR, S. (1999). *Sampling: Design and Analysis*. New York: Duxbury Press, 255-287.

MASSEY, J., and BOTMAN, S. (1988). Weighting adjustments for random digit dialed surveys. In *Telephone Survey Methodology*, (Eds. R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls and J. Waksberg). New York: John Wiley and Sons, 143-160.

SMITH, T. (1990). Phone home? An analysis of household telephone ownership. *International Journal of Public Opinion Research*, 2, 369-390.

Survey Methodology, December 2001
Vol. 27, No. 2, pp. 131-142
Statistics Canada

The Effect of Intensity of Effort to Reach Survey Respondents: A Toronto Smoking Survey

LOUIS T. MARIANO and JOSEPH B. KADANE[1]

ABSTRACT

The number of calls in a telephone survey is used as an indicator of how difficult an intended respondent is to reach. This permits a probabilistic division of the non-respondents into non-susceptibles (those who will always refuse to respond), and the susceptible non-respondents (those who were not available to respond) in a model of the non-response. Further, it permits stochastic estimation of the views of the latter group and an evaluation of whether the non-response is ignorable for inference about the dependent variable. These ideas are implemented on the data from a survey in Metropolian Toronto of attitudes toward smoking in the workplace. Using a Bayesian model, the posterior distribution of the model parameters is sampled by Markov Chain Monte Carlo methods. The results reveal that the non-response is not ignorable and those who do not respond are twice as likely to favor unrestricted smoking in the workplace as are those who do.

KEY WORDS: Call-backs, number of; Bayesian analysis; Markov Chain Monte Carlo method; Informative non-response; Ignorable non-response.

1. INTRODUCTION

Given the reality of non-response in every survey, it is of interest to determine how to account for this non-response in the interpretation of the collected data. Rubin (1976) gives necessary and sufficient conditions for such an analysis to be identical from, respectively, a frequentist, likelihood, and Bayesian perspectives, to an analysis based on a model incorporating a missingness mechanism. Building on this, Little and Rubin (1987) led to an extensive literature modeling non-response in an informative, non-ignorable way.

Information about the interaction between the survey and the surveyed can sharpen the analysis of the import of missing data in a survey. The example in this paper concerns the attitudes of Toronto citizens about smoking in the workplace. Random telephone numbers were chosen; at least twelve calls were made to try to reach the intended respondents. Our data for the respondents includes only the number of calls until the survey was completed, not the timing of the unsuccessful calls. With even this attenuated data on how difficult the respondent was to reach, we find our view of the results of the survey to be importantly informed by the number of unsuccessful calls.

The use of information on the number of calls to a subject chosen to participate in a survey is not unique. Potthoff, Manton and Woodbury (1993) present a method for correcting for survey bias due to non-availability by weighting based on the number of call-backs. While our analysis also focuses on the bias due to non-availability, there are major differences. Instead of assuming that refusals do not exist, we allow for and utilize their potential existence in modeling the mechanism which causes non-response. In the analysis that follows, the relationship of non-response to the response variable of interest in the survey is evaluated along with other explanatory variables, after weighting for both household size and the appropriate population demographics. In doing so we address not only whether error exists due to non-availability, but also whether stratification of the respondents by household size and the then current age/sex distribution may eliminate the necessity for accounting for the error by the introduction of a mechanism which describes the non-response. Note that here we match the groupings of Pederson, Bull and Ashley (1996) used in the original published analyses of the dataset; more complex cell adjustment procedures are possible (*e.g.*, Little 1996; Eltinge and Yansaneh 1997, and references cited therein).

The remainder of this article is organized as follows: Section 2 gives more detail on the survey; section 3 introduces the methodology employed; Sections 4 and 5 respectively explore missing-at-random and non-ignorably-missing models; Section 6 discusses the priors distributions chosen for the main analysis, whose results are explained in section 7. Finally, section 8 gives our conclusions.

2. THE SURVEY

A bylaw regulating smoking in the workplace in the City of Toronto took effect on March 1, 1988. From January 1988 to the present, a series of six surveys have been conducted to assess attitudes of the public toward smoking, awareness of health risks related to smoking, and the impact of the law on the residents of Metropolitan Toronto. The data being utilized in this analysis comprises the third phase

[1] Louis T. Mariano is a Ph.D. candidate, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213; Joseph B. Kadane is Leonard J. Savage University Professor of Statistics and Social Sciences, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213.

of this series. Northrup (1993) provides the technical documentation for this survey. For clarity, when necessary, the data being analyzed here is referred to as the Phase III data, and information from the first two surveys is referred to as the Phase I & II data.

Northrup (1993) indicates that the data of interest, which were made available by the Institute for Social Research (ISR) at York University, were collected from 1,429 residents of the Metropolitan Toronto area in December 1992 and March 1993. A two-stage probability selection process was utilized to select survey respondents. The first stage employed random digit dialing. The second stage used the most recent birthday method to select one adult individual once an eligible residence was reached. The responses were then weighted by the number of adults in the household. In the analysis that follows, post-stratification weighting was also applied to the census age-sex distribution to adjust for the underrepresentation of some population subgroups. The number of distinct phone lines in the household was not taken into consideration during the data collection.

The number of calls it took to reach each respondent is included as a variable in the dataset, and there are no missing values for this variable. Northrup (1993) explains that the 1,429 responses came from a sample of 5,702 telephone numbers generated by the random digit dialing method. Of these numbers, 2,286 were verified to be eligible households, and 3,150 of the numbers in the sample were not eligible. The status of the remaining 266 numbers was not able to be determined. It has been assumed by ISR that the household eligibility rate of these 266 numbers was equal to the rate for the rest of the sample. This eligibility rate implies an estimated total of 2,398 households in the sample and a response rate of 60%. Thus, an estimated 969 subjects chosen to participate in the survey did not respond. Each subject received a minimum of 12 calls, including day, night, and weekend calls, before being classified as non-respondent.

The dependent variable, for the purpose of this analysis, is an individual's opinion on the regulation of smoking in the workplace, in one of three categories. Category "0" indicates smoking should be permitted in restricted areas only, category "1" indicates smoking should not be permitted at all, and category "2" indicates smoking should not be restricted at all. For each subject chosen to participate in the survey, let $Y_i \in \{0, 1, 2\}$ represent the opinion of subject i.

The data comprises of the answers to 50 survey questions as well as 18 other variables identifying characteristics of the subject. Included in these are:

- "K-risk" is an integer score from 0 to 12 which indicates knowledge of the risks and effects of second-hand smoke.
- "Smoker" indicates the smoking status of the subject: "Current smoker" (S), "Former smoker" (SQ) or, "Never smoked" (NS).
- "Bother" indicates if second-hand smoke bothers the subject: "Always bothers" (b.A), "Usually bothers" (b.USUL), or "Does not bother" (b.NO).
- "Age": (Age in years - 50) / 10.

Pederson, Bull, Ashley and Lefcoe (1989) created a "Knowledge of health effects score" on passive smoking out of the answers to six survey questions, which measured a subject's knowledge of the effects of second-hand smoke. Pederson *et al.*'s questions were used in Phase III to create their score, here renamed "K-risk". A higher K-risk score indicates a greater knowledge of the risks of second-hand smoke. The variable "Age" was shifted and rescaled to match how age was treated by Bull (1994) in the Phase I & II analysis.

3. OVERVIEW OF METHODOLOGY

The fundamental question of interest is: "May we ignore the unit non-response and treat the observed data as a random subsample of the population?" Mapping to the terminology of Little and Rubin (1987) and Rubin (1976): If we may treat the observed data for the dependent variable of interest as a random subsample, we call the missing data "missing completely at random" (MCAR). If we may treat the observed data for the dependent variable of interest as a random subsample, after conditioning on the explanatory variables, we call the missing data "missing at random" (MAR). Let θ represent the parameters of the data and let π represent the parameters describing the missing data process. Rubin (1976) calls the parameters π and θ distinct "if there are no *a priori* ties, via parameter space restrictions or prior distributions, between π and θ." If either the MCAR or MAR cases apply and if π and θ are distinct, the mechanism which causes the missing data is said to be "ignorable" for inference about the distribution of the variable of interest. If the missing data for the dependent variable of interest is dependent on the values of that data, then the mechanism which causes the missing data is said to be "non-ignorable" (NI). Groves and Couper (1998) note that when the likelihood of participation is a function of the desired response variable, the non-response bias can be relatively high, even with a good response rate.

Let R_i be an indicator of response. $R_i = I_{\{\text{respondent}\}}(\text{subject } i)$ and $R = (R_1, ..., R_n)^T$. Little and Rubin (1987) suggest that one possible method for accounting for the non-response mechanism is to include this response indicator variable in the model. We may call the mechanism which causes the missing data ignorable if π and θ are distinct and:

$$f(R \mid Y_{obs}, Y_{mis}, \pi) = f(R \mid Y_{obs}, \pi) \qquad (1)$$

where Y_{obs} and Y_{mis} represent the observed and missing portions of the dependent variable of interest.

The terms "MAR assumption" and "NI assumption" will be used throughout this analysis. For clarity, the term "MAR assumption" is defined as the assumption that the missing data mechanism is ignorable for inference with respect to the dependent variable identified in section 2. That is, the observed values of that variable are a random subsample of the population, possibly within poststrata, and it is not necessary to account for the missing data mechanism. The term "NI assumption" is defined as the assumption that the missing data mechanism is non-ignorable and the data collected for the dependent variable of interest cannot be treated as a random subsample. Specifically, inference for the population must involve the missing data mechanism.

The approach to assessing the MAR assumption is comprised of three steps. The first step is the examination of what one might do under the MAR assumption. Since the dependent variable of interest has three categories and some of the explanatory variables are quantitative, polytomous logistic regression is employed. Both frequentist and Bayesian forms of the logistic regression model are examined.

In the second step, and NI model is constructed. The non-response mechanism is modeled utilizing the information available about the number of calls made to each subject. Here, the idea of a surviving fraction in the sample is examined to model whether it is actually possible to reach all the intended respondents. Then, the non-response mechanism is related to the dependent variable by including the number of calls in the logistic regression model.

In the development of the NI model, we employ a Bayesian approach to allow for an examination of the values the missing data re likely to take, given the observed data and the model parameters. This is accomplished by utilizing a data augmentation approach, where the missing data are imputed in each iteration of a Markov Chain Monte Carlo (MCMC) simulation. A possible alternative would be to utilize the expectation-maximization (EM) algorithm (Dempster, Laird and Rubin 1977) to compute the maximum likelihood estimates (MLE's) of the missing values.

In the third step, an evaluation of the MAR assumption is made. Non-zero coefficients for the number of calls in the logistic regression portion of the NI model will imply that the number of calls does make a difference; *i.e.*, the opinions of those who did not respond in the first 12 calls are likely to differ from those who responded in just a small number of calls. In this case, the missing data mechanism is not independent of the values of the missing data and an MAR assumption would be inappropriate. Next, the log odds of response among the three models are examined. Differences here identify the magnitude of the error that a faulty MAR assumption causes. So, in the evaluation of the MAR assumption, the questions "is there a difference?" and "how large is the difference?" are both addressed.

4. MAR MODELS

4.1 Logistic Regression

Using the data collected from the (m = 1,429) subjects that did respond to the survey, weighted logistic regression was employed to model the public's opinion on smoking in the workplace. The collection of candidate predictors found in the survey questions and the background information was narrowed utilizing a series of Wald tests. Then likelihood ratio tests, AIC, and BIC were used to compare the possible models. The model with the best fit was found to be the one which included additive terms for the variables "K-risk", "Smoker", "Bother", and "Age", as defined in section 2.

As each of the models examined in this analysis employs a logistic regression component, it is useful here to illustrate the notation being used. Category "0", "smoking allowed in restricted areas only" was chosen to be the reference category. Recall $Y_i \in \{0, 1, 2\}$. For the MAR model, we use only the observed values of the subject's opinion on workplace smoking, $Y_{obs} = (Y_1, ..., Y_m)$. Let $Y_{ij} = I_{\{j\}}(Y_i)$ be an indicator of subject i responding in category j, and let W_i represent the weight each subject received. As in the original published analyses of this dataset (Pederson *et al.* 1996) both household (see Northrup 1993) and post-stratification (see Appendix A) weighting were used in the consideration of all models here.

The two categorical explanatory variables, "Smoker" and "Bother", were included in the model by utilizing indicator variables for two of the three categories, with the effect of the third category being absorbed in the intercept term. For "Smoker", "S_i" and "SQ_i" were included as indicators that subject i was either a current smoker or a smoker who had quit. For "Bother", "$b.USUL_i$" and "$b.NO_i$"were included as indicators that second had smoke usually bothered or did not bother subject i.

Let X_i = represent the vector for explanatory variables for subject i. Then,

$$X_i = (\text{K-risk}_i, S_i, SQ_i, b.USUL_i, b.NO_i, \text{Age}_i).$$

Here we use an unordered multinomial logit model to consider $p_j(x_i) = P(Y_{ij} = 1 | X_i = x_i)$, the probability that subject i responds in category $j \in \{0, 1, 2\}$, given the observed explanatory variables for subject i. This model, of course, utilizes linear equations η_{ij} describing the log odds of subject i responding in category j versus the reference category $j = 0$. So, for $j = 1, 2$ we wish to examine:

$$ln\frac{p_j(x_i)}{p_0(x_i)} = \eta_{ij} = \beta_{0j} + X_i\beta_j, \qquad (2)$$

with $\eta_{i0} = 0$. The two resultant linear equations, η_{i1} and η_{i2}, each have seven coefficients, including an intercept term β_{0j} and those displayed below:

$$\beta_j = (\beta_{\text{K-risk}_j}, \beta_{S_j}, \beta_{SQ_j}, B_{b.USUL_j}, \beta_{b.NO_j}, \beta_{\text{Age}_j}).$$

The MAR logistic regression model has 14 parameters. The vector of these 14 parameters, represented by $\beta = (\beta_{01}, \beta_1, \beta_{02}, \beta_2)$ has the likelihood (or, more appropriately, pseudo-likelihood, since the weights are incorporated through the variable W_i):

$$L(\beta) \propto \prod_{i=1}^{m} \prod_{j=0}^{2} \left(\frac{e^{\eta_{ij}}}{1 + e^{\eta_{i1}} + e^{\eta_{i2}}} \right)^{y_{ij} w_i} . \quad (3)$$

4.2 Bayesian Logistic Regression

The likelihood in equation (3) and the data collected from the survey respondents are utilized in the Bayesian analysis. The same four explanatory variables selected in the frequentist analysis above are used as the explanatory variables here. Prior distributions, discussed in section 6, were assigned to the logistic regression parameters. An MCMC simulation is utilized in order to draw from the posterior distribution of the parameters.

5. NI MODEL

5.1 Modeling the Non-Response Mechanism

Since the missing values are not necessarily missing at random, the mechanism which caused them to be missing must be addressed. Northrup (1993) indicates that non-respondent subjects chosen to participate in the survey were called a minimum of 12 times, including a minimum of three day, four evening and four weekend calls. Unfortunately, other useful information regarding the number of calls was not retained. We do not know which of the non-respondents were called more than twelve times or whether an individual call was placed during the day, evening, or weekend. We also are unaware of the details of the non-response, such as whether the subject was contacted but refused to participate, whether the calls were ever answered by a machine, or whether they were answered at all. Thus, stratification of the non-respondents was not possible, and they were all treated as exchangeable in this analysis.

Each subject was called a number of times until the survey was successfully completed or they were classified as non-respondent. For the respondents, the number of calls variable (C_i) describes the number of trials until the first success for subject i. Thus, one might expect the number of calls to follow a Geometric distribution with truncated observations for the non-respondents. Specifically, let $\pi = P(\text{a call is successful})$; then, consider $C_i \sim \text{Geometric}(\pi)$ and $P(C_i = c_i) = \pi(1-\pi)^{c_i - 1}$. Note that if auxiliary information about the number of calls to the non-respondents were available (*e.g.*, Groves and Couper 1998), we could have also considered conditional response probabilities here.

The histograms in Figure 1 compare the data (through the first twelve calls) to a Geometric distribution with parameter $\pi = .225$, which appears to match fairly well. The sample order statistics suggest $\pi \in (.2, .25)$. The histogram of the actual survey data reveals that the number of subjects reached on the first call are fewer than the number reached on the second call. It is possible that more of the second calls were placed at a time which had a higher success rate.

Suppose $\pi = .225$; by the memoryless property of the Geometric distribution, we would expect 218 of the 969 non-respondents to reply on the 13th call. This would make the data through the first 13 calls appear as in Figure 2. Clearly, Figure 2 does not display the behavior of a Geometric random variable. Consider the following question: "If all subjects were called an unlimited amount of times, would they all have been reached?" Answering "yes" to that question for this dataset results in the problem illustrated in Figure 2.

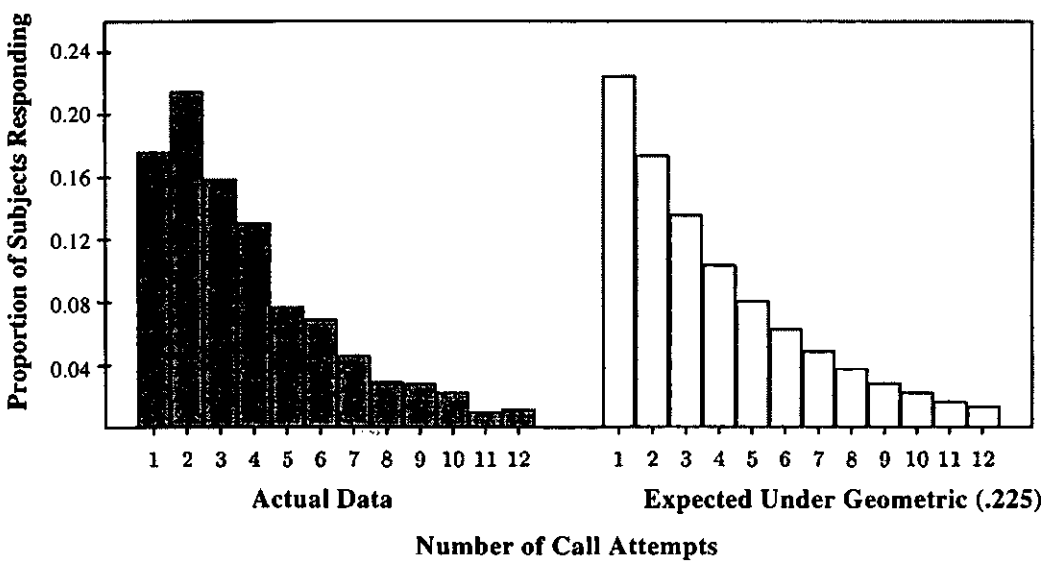

Figure 1. Comparison of the actual survey data for sucessful calls in the first 12 attempts to expected results based on a Geometric (.225) distribution for the number of calls needed to complete the survey.

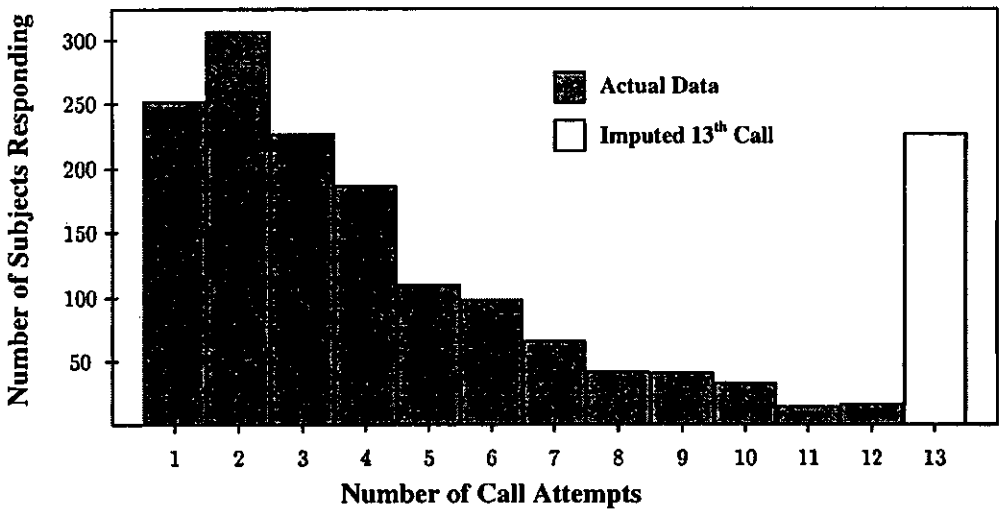

Figure 2. Display of the actual number of successful calls on each attempts through the first 12 and the expected number of successful calls on the 13th attempt. The expectation for the 13th callis based on a Geometric (.225) distribution to model the number of calls until the survey is completed

Given the information outlined above, the assertion that "not all subjects chosen for the survey are reachable" is a viable one. Maller and Zhou (1996) discuss immune subjects – individuals who are not subject to the event of interest. Following their terminology, if it is not possible to procure a response from a subject chosen for the survey given an unlimited amount of calls, that subject is categorized as immune. Subjects who are not immune are categorized as "susceptible". The set of immune (*i.e.*, non-susceptible) subjects comprise the "surviving fraction" of the sample. Mapping to more familiar terminology, the immune subjects include those who were reached and refused, those who would have refused if they had been reached, and those cases of a physical or mental inability to ever participate. Northrup (1993) indicates that those who initially refused to participate were subsequently contacted by the most senior interviewers, so, we make the assumption here that all remaining refusals would not ever participate. The susceptible group includes the respondents, those who would have responded if successfully contacted, and those who were physically or mentally unable to participate during the data collection period but were willing and able at some other time.

Let the variable $Z_i = I_{\{\text{susceptible}\}}(\text{subject } i)$ be an indicator of the susceptibility of subject i, and $\rho = P(\text{subject } i \text{ is susceptible})$, *i.e.*, $Z_i \sim \text{Bernoulli}(\rho)$. Now suppose that the number of calls to the susceptible subjects follows a Geometric distribution, *i.e.*, $C_i \mid Z_i = 1 \sim \text{Geometric}(\pi)$. Does this eliminate the problem illustrated in Figure 2?

Let R_i be an indicator of response of subject i. The non-response mechanism can be accounted for by including these response indicators in the model. However, the introduction of the susceptibility variable implies two distinct classes of non-response. So, it is possible to be more detailed and use both the susceptibility $Z = (Z_1, ..., Z_n)^T$ and the response R indicators in a mixture model describing the non-response. Updating Equation (1), the missing data mechanism is ignorable if and only if (π, ρ) is distinct from θ and

$$f(R, Z \mid Y_{obs}, Y_{mis}, \pi, \rho) = f(R, Z \mid Y_{obs}, \pi, \rho). \quad (4)$$

Let $C_{obs} = (C_1, ..., C_m)$ and $Z_{obs} = (Z_1, ..., Z_m)$ be the vectors of the number of calls and the observed susceptibility for each respondent. Also, let $R = (R_1, ..., R_n) =$ be the vector of response for each intended respondent. Every subject, i, may be classified by response into three mutually exclusive groups, A_{obs} - observed, A_{mis} - missing, and A_{imm} - immune, where:

A_{obs} = {i: i was Susceptible and Responded}

A_{mis} = {i: i was Susceptible but did not Respond in 12 calls}

A_{imm} = {i: i was not Susceptible}.

The probability that a subject is in each of these categories may be calculated as follows:

$$P(i \in A_{obs}) = P(Z_i = 1, R_i = 1, C_i = c_i) = \rho\pi(1-\pi)^{c_i - 1}$$

$$P(i \in A_{mis}) = P(Z_i = 1, R_i = 0, C_i > 12) = \rho(1-\pi)^{12}$$

$$P(i \in A_{imm}) = P(Z_i = 0) = 1 - \rho.$$

The data indicates m = 1,429 subjects in A_{obs} and $n - m$ = 969 non-responsive subjects in $A_{mis} \cup A_{imm}$; n=2,398 is the estimated total number of subjects chosen to participate in the survey. Thus, the joint density of Z_{obs}, R and C_{obs} given ρ and π is:

$$f(Z_{obs}, R, C_{obs} \mid \rho, \pi) \propto$$

$$\left[\rho^m \pi^m (1-\pi)^{(\sum_{i=1}^{m} c_i)-m}\right] \times \left[(1-\rho) + \rho(1-\pi)^{12}\right]^{n-m} . \quad (5)$$

The mixture model described by Equation 5 may be viewed as a special case of the non-response models discussed in Drew and Fuller (1981).

It would be useful to confirm that the above joint distribution accurately represents the response pattern of the susceptibles in the dataset. The MLE estimate for ρ is simply the proportion of respondents in the sample, which clearly underestimates ρ. Setting $U(0, 1)$ prior distributions for both ρ and π and examining their joint posterior distribution by MCMC simulation, the posterior medians are found to be ρ =.636 and π =.205, with equal-tailed posterior credible intervals of (.613, .659) and (.191, .219) for ρ and π respectively. Figure 3 illustrates how the dataset might look after imputing the missing number of calls for our susceptible non-respondents based on these posterior medians. The problem previously displayed in Figure 2 has now been mostly eliminated.

While the Geometric distribution appears sufficient (after accounting for susceptibility), a referee questions the use of the Geometric distribution as it does not make use of possibly useful covariates. As explained above, the covariates we think would be most useful for this purpose were not collected. One alternative for modeling the response mechanism of the susceptibles is to use a discretized Gamma distribution. In cases where more complexity is necessary, the v-Poisson (a two parameter Poisson which generalizes some well known discrete distributions, including the Geometric) of Shmueli, Minka, Kadane, Borle and Boatwright (2001) may also be considered.

5.2 Relating Non-Response to the Dependent Variable – The NI Model

Since the non-response of the susceptibles is described by the conditional Geometric distribution of the number of calls, the effect of the non-response of the susceptibles on the dependent variable may be considered by including the number of calls as an additional explanatory variable in the logistic regression likelihood. This will create two additional parameters in the logistic regression portion of the model, which are the coefficients of the number of calls, β_{call} in each of the linear equations η_{ij} described in equation (2).

Non-zero coefficients for the number of calls, then, would indicate that the dependent variable is not independent of the non-response mechanism, and, hence the non-response mechanism is non-ignorable. If these coefficients are zero, the non-response of the susceptibles is ignorable. Conclusions made here rely upon the underlying modeling assumption that the relationship among the number of calls, the dependent variable and the other explanatory variables considered is the same for the respondents and susceptible non-respondents. Including the number of calls in the logistic regression portion of the model does not address the immune subjects, since there will never be the realization of a successful call to them.

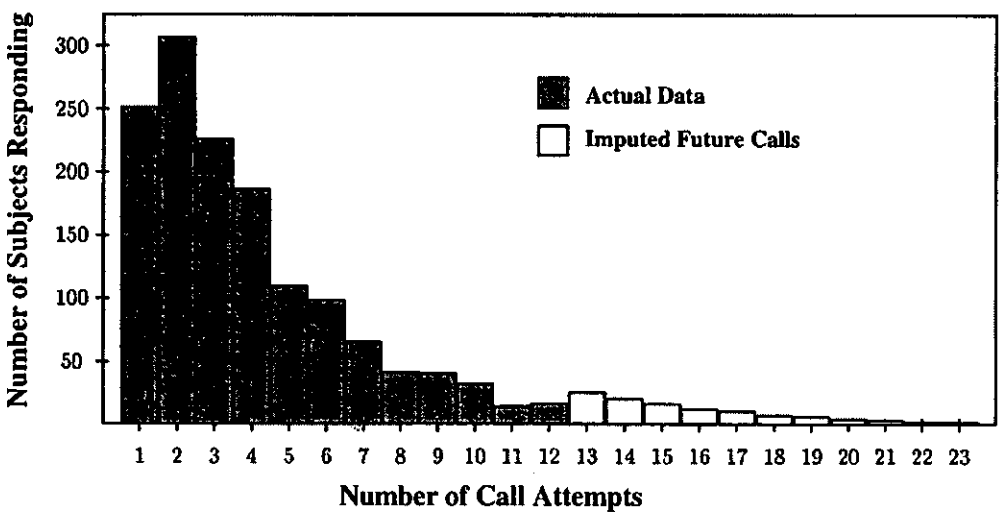

Figure 3. Display of the actual number of successful calls on each attempt through the first 12 and the expected number of successful calls for call attempts 13 and higher. Imputed values are based on a probability of a successful call of .205 and a probability of susceptibility of .636.

The full pseudo-likelihood for the NI model (or, more precisely, the susceptible NI model) is the product of the non-response and logistic regression pieces:

$$L(\rho,\pi,\beta) \propto$$

$$\left[\rho^m \pi^m (1-\pi)^{(\sum_{i=1}^{m} c_i)-m}\right] \times \left[(1-\rho)+\rho(1-\pi)^{12}\right]^{n-m}$$

$$\times \left[\prod_{i=1}^{m}\prod_{j=0}^{2}\left(\frac{e^{\eta_{ij}}}{1+e^{\eta_{i1}}+e^{\eta_{i2}}}\right)^{y_{ij}w_i}\right]. \quad (6)$$

Note that the household and post-stratification weighting variable W_i is included here in an effort to account for whether proper stratification of the respondents may eliminate the necessity for the introduction of a mechanism to describe non-response.

5.3 Data Augmentation

Tanner and Wong (1987) suggest an iterative method for computation of posterior distributions when faced with missing data. This method applies whenever augmenting the dataset makes it easier to analyze and the augmented items are easily generated. Consider the following additional notation: Let S represent the total number of susceptible subjects in the sample. $S = \sum_{i=1}^{n} Z_i$, $S \sim$ Binomial(ρ). Let X be the matrix of explanatory variables (including the number of calls) for all the subjects selected to participate in the survey. Let $Y = (Y_1, \ldots, Y_n)$ be the vector of their responses. Partitions X into $\{X_{obs}, X_{mis}, X_{imm}\}$ and Y into $\{Y_{obs}, Y_{mis}, Y_{imm}\}$. Also, by the memoryless property of the Geometric distribution, the distribution of the additional number of calls required to reach the subjects in A_{mis} is known, and may be expressed: $\forall i \in A_{mis}$, let $V_i = C_i - 12$, which is also distributed as a Geometric random variable with parameter π.

Now suppose that the true values of S, X_{mis}, and Y_{mis} were known. The likelihood could then be considered in the form:

$$L(\rho,\pi,\beta \mid X_{obs}, X_{mis}, Y_{obs}, Y_{mis}, S, R)$$

$$\propto \left[(\rho\pi)^s (1-\pi)^{(\sum c_{sus})-s}\right] \times \left[(1-\rho)^{n-s}\right]$$

$$\times \left[\prod_{i=1}^{s}\prod_{j=0}^{2}\left(\frac{e^{\eta_{ij}}}{1+e^{\eta_{i1}}+e^{\eta_{i2}}}\right)^{y_{ij}w_i}\right], \quad (7)$$

where $\sum C_{sus} = \sum C_{obs} + \sum (V_i + 12)$ is the number of calls that would have been necessary to reach all susceptibles and the summands are taken over the appropriate range of subjects.

Although the true values of S, X_{mis}, and Y_{mis} are unknown, one may utilize what is known about the behavior of these variables to impute stochastically possible values for them within the MCMC algorithm. Given ρ, a value for S may be drawn from a truncated Binomial (2,398, ρ), where $1{,}429 \le S \le 2{,}398$. Given S, the number of subjects in A_{mis} is known. For each of these subjects in A_{mis} a value $V_i \sim$ Geometric (π) may be drawn, which results in an imputation for the number of calls needed to reach each susceptible but unreached subject. The relationships among the number of calls and the other explanatory variables may then be exploited to impute values for the rest of X_{mis}. Specifically, the missing values of Age and K-risk are imputed by regressing Calls on Age and K-risk respectively and predicting from the resultant linear equations. Similarly, the missing values of Smoker and Bother are imputed via logistic regression on each, using Calls as the explanatory variable. Here the model assumptions are checked using the respondents data, and an assumption is being made that these same relationships hold for the susceptible non-respondents. Note that these regression and logistic regression equations are fit in the Bayesian context (*e.g.*, Gelman, Carlin, Stern and Rubin 1998) and necessitate the inclusion of additional parameters, β_I, in the MCMC process which describe these relationships (see Appendix B for more detail). We chose this imputation plan in the interest of the efficiency of the full MCMC algorithm. An alternative would be to impute the missing values for a particular explanatory variable conditional on all the remaining variables (*e.g.*, Rubin 1996). Finally, Y_{mis} may be predicted by utilizing the imputed values of X_{mis} and the relationship described in the logistic regression model. In the interest of the exchangeability of the susceptible non-respondents in the absence of subsequent stratification information, we apply a weight of 1.0 to all the imputed Y_{mis} values; an alternative here would be to impute the sex and household size of the susceptible non-respondents, in addition to their age, and apply the weighting procedure described in Appendix A to the imputed Y_{mis}.

5.4 Sampling from the Posterior Distribution

The full MCMC simulation consists of a Metropolis algorithm supplemented in every iteration with the data augmentation described above. An outline of the MCMC algorithm used may be found in Appendix B. Convergence was assessed utilizing the method of Hiedelberger and Welch (1983) as described in Cowles and Carlin (1996). MacEachern and Berliner (1994) assert that, under loose conditions, subsampling the MCMC simulated values to account for autocorrelation will result in poorer estimators. Following their suggestion, all simulated values, after an appropriate burn-in period, were used in the analysis that follows.

6. CHOICE OF PRIOR DISTRIBUTIONS

In the evaluation of possible prior distributions for the parameters of both the NI and MAR models, the goal of the comparison of the various models was taken into consideration. The choice of prior distributions for the parameters was made from the perspective of the MAR belief. Two possibilities were examined.

The first option is built around the utilization of the Phase I & II surveys. Since these surveys were similar to and were completed prior to the Phase III survey which comprises our data, information contained in these first two surveys may be utilized in the construction of priors. The same dependent variable was contained in the Phase I & II dataset, along with the variables Smoker, Age, and K-risk. A logistic regression model was compiled from the Phase I & II data to describe the relationship between the opinion on workplace smoking and these three explanatory variables. Normal priors were constructed for the coefficients of these three variables centered at their MLE's, but with increased standard error. The error terms were increased due to three factors:

i) There was a three year span between the Phase II and Phase III surveys; opinions may have changed over that time, possibly as a result of the impact of the bylaw.

ii) The MLE's were calculated under the same MAR assumption being evaluated.

iii) Prior to the collection of the Phase III data, there existed the possibility that other explanatory variables would be included in the model; in the presence of other variables, the effect of these three could be altered.

Although the variances were increased, the means were not changed, since it was unknown, *a priori*, in what direction any change might occur. Since the available Phase I & II data contained no information about the Calls or Bother variables, the coefficients of these were assigned a diffuse Normal (0,9) prior. For clarity, this option will be referred to as the "Phase I & II prior" in this analysis.

In the second option Normal (0,9) priors are assigned to each of the logistic regression coefficients. One motivation for this choice is that, for the same three reasons the error terms were increased above, the variables common to the Phase I & II and Phase III surveys are not exchangeable. Thus, construction based on the Phase I & II results would be inappropriate. This option will be referred to as the "Central prior".

The choice to use Normal (0,9) distributions here is for convenience. Centering the prior at zero gives equal weight to either direction of the relationship. We believe the choice of a variance of nine to be adequate without being overly diffuse. The use of improper priors could lead to a Markov Chain Monte Carlo simulation that never converges; and, as Natarajan and Kass (2000) show, an overly diffuse proper prior may behave like an improper one. In section (7.2), we offer a sensitivity analysis to evaluate how the results are effected by the choice of prior.

The non-response parameters of the NI model, ρ and π, were treated the same under both prior options. There was no additional information available about the probability of a successful call or the probability of susceptibility. Thus, ρ and π were each assigned a $U(0,1)$ prior.

The data augmentation parameters found in each of the logistic regression equations, β_l, were independently given diffuse Normal (0,9) priors. For each linear regression equation found in the data augmentation process, the coefficients, β_r, and variance, σ_r^2, were set to $p(\beta_r, \sigma_r^2) \propto 1/\sigma_r^2$, the standard non-informative prior distribution (*e.g.*, Gelman *et al.* 1998). Note that the closed forms of the posterior distributions of the linear regression parameters are known and may be drawn from directly.

7. RESULTS

First, the validity of the MAR assumption is examined through the coefficients of the number of calls variable. Then, the NI model is evaluated with respect to sensitivity to the choice of prior. Finally, the magnitude of the impact of a faulty MAR assumption for this dataset is investigated by illustrating the change in the odds of response.

7.1 Coefficients for the Number of Calls

For both the Phase I & II and Central priors, Figure 4 displays the posterior density (solid line) and 95% credible interval estimates (dotted lines) of the coefficient of the calls variable in η_{i1} in the NI model, and compares them to the point $\beta_{call_1} = 0$ (dashed lines). The results clearly indicate this coefficient differs from zero. We also find a non-zero result in η_{i2}, where, using the Phase I & II prior, the 95% HPD credible interval for β_{call_2} is (-0.03613, 0.11595).

The non-zero coefficient of C_i demonstrates a dependence between the number of calls and the subject's opinion on smoking in the workplace. Thus, the dependent variable and the non-response mechanism are not independent under the conditions discussed in section 5.2. This results implies that an assumption that the missing observations are missing at random prior to accounting for the non-response mechanism is incorrect for this dataset.

There is a hint in Figure 3 that the probability of a successful call decreases as the call number increases. To verify the assumption that the relationship between the number of calls and the log odds of response is linear, a second Bayesian NI model was constructed. This model split the calls variable into two, $C_i I_{\{C_i<7\}}$ and $C_i I_{\{C_i \geq 7\}}$, based on whether the number of calls were fewer than seven. The posterior distributions of the coefficients of these two variables were then compared and evidence that they are essentially different was not found. In particular, for η_{i1} the 95% credible interval for $C_i I_{\{C_i \geq 7\}}$ contained the same interval for $C_i I_{\{C_i<7\}}$, and for η_{i2} the 95% credible intervals strongly overlapped.

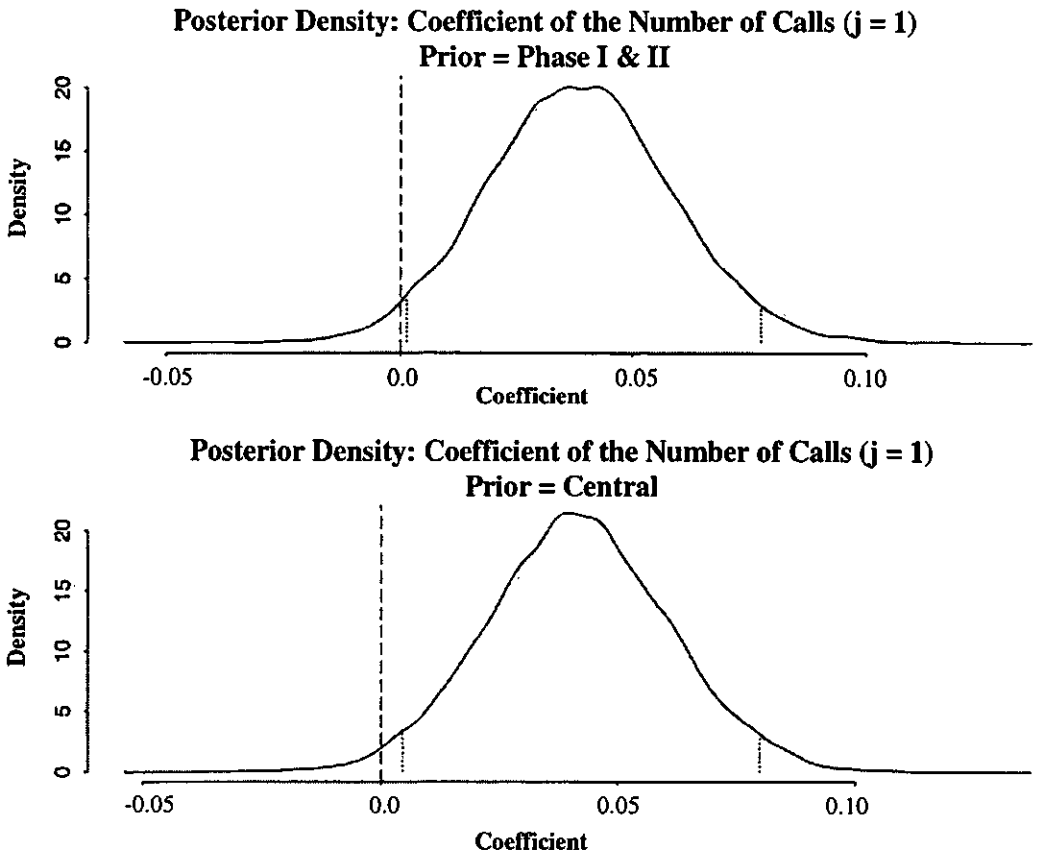

Figure 4. Display of β_{call_1}, the coefficient of the calls variable in η_{i1}: posterior density (solid line) and 95% equal tailed credible interval (dotted line), compared to $\beta_{call_1} = 0$ (dashed line).

7.2 Sensitivity to Priors

Would different prior distributions, either on the calls coefficient or on the others, make a difference in the effect illustrated above? Table 1 displays 95% HPD credible intervals for the coefficient of the calls variable in the first logit equation of the NI model for six different priors. The priors include the Phase I & II and Central priors as well as four others - labeled options 3, 4, 5, and 6. Options 3 and 4 resemble the Central prior except that they change the prior distribution on the coefficient of the number of calls to Normal (1,9) and Normal (-1,9) respectively. Option 5 places Normal (0,9) priors on β_{call_1}, β_{age_1}, and $\beta_{b.USUL_1}$, a Normal (1,9) prior on β_{01}, a Normal (.5,9) prior on $\beta_{K\text{-}risk_1}$, a Normal (-1,9) prior on β_{S_1} and Normal (-5.9) priors on β_{SQ_1} and $\beta_{b.NO_1}$. Option 6 takes the Central Prior and reduces all the variances from nine to two.

Under all six priors, Table 1 demonstrates that the coefficient of the calls variable in the first logit equation clearly differs from zero. The finding that the missing data mechanism is non-ignorable for this dataset does not appear to be effected by the choice of prior among these options.

Table 1
95% HPD Credible Intervals for β_{call_1} Under six Different Prior Distributions

Prior	Coefficient of the number of Calls "C_i" in η_{i1}	
	95% intervals	
	Lower Bound	Upper Bound
Phase I & II	0.00129	0.07746
Central	0.00446	0.07980
Option 3	0.00447	0.07983
Option 4	0.00441	0.07975
Option 5	0.00440	0.07970
Option 6	0.00436	0.07944

7.3 Effect on Odds of Response

Given the failure of the MAR assumption shown above, it is of interest to question the relevance of the error that using the MAR assumption would create. The magnitude of the error induced by a faulty MAR assumption may be illustrated by examination of its effect on the odds ratio $p_1(x_i)/p_0(x_i)$. First, we consider the effect on a typical

respondent profile. The modal respondent was a non-smoker between the ages of 25-35 years old who was usually bothered by second-hand smoke, had a K-risk of 11 and could be reached in 2 calls. We label this modal respondent as Subject 1. Table 2 demonstrates the change in posterior odds for Subject 1 when called 13 times.

Table 2
Comparison of the Odds of Response for 4 Typical Subjects. Posterior Medians Were Used As the Point Estimates for the Coefficients in the Bayesian Models; the Mle Was Used for the Frequentist Model

	Subject 1	Subject 2	Subject 3	Subject 4
Smoker	No	No	Former	Yes
Age	30	50	27	40
Bother	Usually	Always	No	No
K-risk	11	12	7	3
Model		Odds Y=1/Y=0		
MAR MLE	0.674	2.105	0.457	0.396
MAR Phase I & II prior	0.703	4.487	0.209	0.116
NI Phase I & II prior: 2 calls	0.640	4.024	0.202	0.108
NI Central prior: 2 calls	0.593	4.442	0.162	0.102
Option 3: 2 calls	0.594	4.449	0.162	0.102
Option 4: 2 calls	0.592	4.435	0.162	0.101
Option 5: 2 calls	0.590	4.423	0.161	0.101
Option 6: 2 calls	0.590	4.426	0.161	0.101
NI Phase I & II prior: 13 calls	0.974	6.128	0.308	0.165
NI Central prior: 13 calls	0.936	7.013	0.256	0.160
Option 3: 13 calls	0.937	7.026	0.256	0.161
Option 4: 13 calls	0.934	7.000	0.255	0.160
Option 5: 13 calls	0.930	6.975	0.254	0.159
Option 6: 13 calls	0.931	6.980	0.254	0.160

The Subject 1 column Table 2 indicates a dramatic difference in the posterior odds when the non-response mechanism is taken into consideration. For this typical respondent profile, when the number of calls is increased from two to thirteen the posterior odds of choosing "Smoking should not be permitted at all" over "Smoking should be permitted in restricted areas only" increases by 52.18% under the Phase I & II prior and 57.84% when using the Central prior. This is dramatic evidence of the relationship between the dependent variable and the non-response mechanism.

Are the results for the modal subject above typical? Table 2 also displays the effects on the odds of response under the NI model for three additional test subject profiles for each of the six different priors considered above. Subject 2 is a fifty year old non-smoker who is always bothered by smoke and has a perfect "K-risk" score. Subject 3 is a 27 year old former smoker who is not bothered by smoke and has a "K-risk" score of seven. Subject 4 is a 40 year old smoker who is not bothered by smoke and has a "K-risk" score of three. On multiple subjects with multiple priors, Table 2 consistently shows the same result. Increasing the number of calls to greater than 12 will increase the posterior odds of choosing category "1" over category "0". For each of the test subjects and priors found in Table 2, the increase was between 52.18% and 58.41%.

Similar results were found when examining the odds of choosing the "Smoking should not be restricted at all" category over the "Smoking should be permitted in restricted areas only" category. Using test subjects which were a current and a former smoker (Subjects 3 and 4 above), the posterior odds increased 46.7% when the number of calls was increased from 2 to 13 under the Phase I & II prior.

7.4 Effect on Probability of Response

With the shift in posterior odds illustrated above comes a corresponding shift in the estimated probabilities that a subject will respond in a particular category. Among the respondents, 57.45% chose category "0", 40.64% chose category "1", and 1.91% chose category "2". The number of non-respondent susceptibles have a posterior median of 469, with a 95% credible interval of (25, 944). On average, 55.88% of the simulated non-respondent susceptibles chose category "0", 40.03% chose category "1", and 4.08% chose category "2". While, for categories "0" and "1", the average values for the non-respondent susceptibles do fall within the 95% confidence intervals for the proportions of the respondents in these categories, the point estimates for each category shift when the non-response mechanism is included in the model. In comparing the category "2" results, we estimate that non-respondents are twice as likely to favor no restrictions on smoking (category "2") than are respondents. While the low number of subjects found in category "2" are unlike to provoke a change in workplace smoking law, the increasely noted in the non-respondents in this category serves as an example of how the lack of proper consideration of the non-respondents could lead to flawed conclusions about the data.

8. CONCLUSION

Section 7 demonstrates that, for the dependent variable of interest in this dataset, an assertion that the missing observations are missing at random, prior to accounting for the missing data mechanism, is incorrect, assuming the relationship among the relevant variables is the same for all susceptible subjects. Furthermore, the use of a faulty MAR assumption in the evaluation of this dependent variable risks serious error in the calculation of the posterior odds and in any conclusion drawn from them. In order to perform a proper evaluation of the opinion on smoking in the workplace in Toronto in early 1993 via the dependent variable of interest in this survey, it is necessary to account for the non-response mechanism in the model structure.

In this analysis, only one simple piece of information, the number of calls, was utilized. A more complete treatment could have been made, had more information been available. Knowledge of the exact number of calls to the non-respondents, instead of a minimum, and the time of day of the calls could have enabled this analysis to be more precise. In addition, knowledge of the type of non-response, refusal or non-availability, and the number of times the non-respondents were actually contacted could have allowed for better classification of the non-respondents. Groves and Couper (1998) point out that statistical errors arising from non-availability and those arising from refusals are likely to differ. As they further comment, the evaluation of how efforts to seek cooperation effect measurement error is an important area of research.

The results illustrated above apply only to this one dependent variable assessing smoking in the workplace in this one dataset. Given the perception that smoking has become less socially acceptable over recent years, it would be reasonable to think that non-response error due to questions about smoking may be more severe than other topics. A comparison of non-response bias including various smoking related questions and others which do not concern smoking may be found in Biemer (2001); this comparison lends no credence to the idea that non-response error is unique to questions relating to smoking.

Although the above results make no implications about the missing data mechanisms in other surveys, there is a clear demonstration here that blindly assuming that the respondents of a survey constitute a random subsample of the population for the variables of interest can be an unwise choice. Information, available at the time of data collection, can enable the evaluation of whether or not the mechanism which causes the non-response is ignorable. In light of this observation, then, it should be of interest to those who work with such data to make use of the available information pertaining to the non-response in the evaluation of that data and to make such information available to others who utilize the dataset. As a general matter, we believe that the collection and analysis of data on where and how respondents were found, as well as how difficult they were to find, is an important future direction for survey methodology and practice.

ACKNOWLEDGEMENTS

This research was funded by National Science Foundation Grant DMS-9801401. The authors thank Shelley Bull for her many helpful comments and suggestions and for assistance in the acquisition of the data and John Eltinge and the anonymous referees and Associate Editor for their valuable comments.

Data from the Attitudes Toward Smoking Legislation, which was funded by Health and Welfare Canada, were made available by the Institute for Social Research at York University. The data were collected by the Institute for Social Research for Dr. Linda Pederson of the University of Western Ontario, Dr. Shelley Bull of the University of Toronto and Dr. Mary Jane Ashley of the University of Toronto. The principal investigators, the Ontario Ministry of Health and the Institute for Social Research bear no responsibility for the analyses and interpretations presented here.

A. Post-stratification Weighting

HHW_i is the household weight of subject i as described in Northrup (1993).

- Let m = the number of respondents.
- Let r = the cumulative number of adults in the responding households.
- Let h_i = the number of adults in subject i's household.
- $HHW_i = h_i \cdot m/r$.

Proportions in the sample falling into the following age groups were calculated for both male and female respondents: 18-24 years, 25-44 years, 45-64 years, and over 65 years old. These proportions were then compared to the age/sex distribution in Metropolitan Toronto.

- Let p_{1i} = the proportion of adult Metropolitan Toronto residents falling into the same age/sex category as subject i, as per the 1991 Census.
- Let p_{2i} = the proportion of survey respondents with the same age and sex categories as subject i.
- $W_i = HHW_i \cdot p_{1i}/p_{2i}$, where W_i is the final post-stratification weight used in the analysis.

B. MCMC Implementation

The full MCMC simulation for the NI model consists of a Metropolis algorithm supplemented with the data augmentation described in section 5.3. The following is an overview of the MCMC algorithm. Variables used below are defined in section 5. At each iteration t,

1. Draw ρ_t for $Beta(s_{t-1} + 1{,}2398 - s_{t-1} + 1)$.
2. Impute s_t from $Binomial\ (\rho_t) \geq 1{,}429$.
3. Impute C_{mis_t}: draw $(s_t - 1{,}429)$ v_i's from $Geometric\ (\pi_{t-1})$ and $\forall c_i \in c_{mis}$, $c_i = v_i + 12$.
4. Draw π_t from $Beta(s_t + 1, \sum c_{sus_t} - s_t + 1)$.
5. Impute values for the rest of X_{mis} by utilizing the relationships with the number of calls, as described in section 5.3

6. Update the additional parameters used in the data augmentation of X_{mis}.
   - Update linear regression parameters, β_r and σ_r by drawing directly from the closed form of their posteriors.
   - Update logistic regression parameters, β_l using a Metropolis step on each.
7. Impute $Y_{\text{mis}_i}: \forall y_i \in y_{\text{mis}}$ draw y_i from a *Multinomial* $(p_0(x_i), p_1(x_i), p_2(x_i))$.
8. Update each β_{kj} using a Metropolis step on the conditional likelihood and a Normal jump function.

REFERENCES

BIEMER, P.P. (2001). Nonresponse bias and measurement bias in a comparison of face to face and telephone interviewing. *Journal of Official Statistics*, 17, 2, 295-320.

BULL, S. (1994). *Case Studies in Biometry*. Analysis of Attitudes Toward Workplace Smoking Restrictions, chapter 16, New York: Wiley and Sons, 249-270.

COWLES, M.K., and CARLIN, B.P. (1996). Markov Chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91, 883-904.

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

DREW, J.H., and FULLER, W.A. (1981). Nonresponse in complex multiphase surveys. *Proceedings of the section on Survey Research Methods, American Statistical Association*, Alexandria, VA, 623-628.

ELTINGE, J.L., and YANSANEH, I.S. (1997). Diagnosis for formation of nonresponse adjustment cells, with and application to income nonresponse in the U.S. Consumer Expenditure Survey. *Survey Methodology*, 23, 33-40.

GELMAN, A., CARLIN, J.B., STERN, H.S. and RUBIN, D.B. (1998). *Bayesian Data Analysis*. Chapter 14, Generalized Linear Models. London: Chapman & Hall.

GROVES, R.M., and COUPER, M.P. (1998). *Nonresponse in Household Interview Surveys*. New York: Wiley and Sons.

HIEDELBERGER, P., and WELCH, P. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, 31, 1109-1144.

LITTLE, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 54, 139-157.

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

MacEACHERN, S.N., and BERLINER, L.M. (1994). Subsampling the Gibbs Sampler. *The American Statistician*, 48, 188-189.

MALLER, R., and ZHOU, X. (1996). *Survival Analysis with Long Term Survivors*. Chichester, UK: Wiley and Sons.

NATARAJAN, R., and KASS, R.E. (2000). Reference Bayesian methods for generalized linear mixed models. *Journal of the American Statistical Association*, 95, 227-237.

NORTHRUP, D.A. (1993). Attitudes Towards Workplace Smoking Legislation: A Survey of Residents of Metropolitan Toronto, Phase III, 1992/93 Technical Documentation. Tech. Rep. Institute for Social Research, York University, Unpublished.

PEDERSON, L.L., BULL, S.B. and ASHLEY, M.J. (1996). Smoking in the workplace: Do smoking patterns and attitudes reflect the legislative environment? *Tobacco Control*, 5, 39-45.

PEDERSON, L.L., BULL, S.B., ASHLEY, M.J. and LEFCOE, N.M. (1989). A population survey on legislative measures to restrict smoking in Ontario: 3. Variables related to attitudes of smokers and nonsmokers. *American Journal of Preventive Medicine*, 5, 313-322.

POTTOFF, R.F., MANTON, K.G. and WOODBURY, M.A. (1993). Correcting for nonavilability bias in surveys by weighting based on number of callbacks. *Journal of the American Statistical Association*, 88, 1197-1207.

RUBIN, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

RUBIN, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 473-489.

SHMUELI, G., MINKA, T.P., KADANE, J.B., BORLE, S. and BOATWRIGHT, P. (2001). Using Computational and Mathematical Methods to Explore a New Distribution: The v-Poisson. Technical Report 740, Department of Statistics Carnegie Mellon University.

TANNER, M.A. and WONG, W.H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82, 528-549.

Survey Methodology, December 2001
Vol. 27, No. 2, pp. 143-154
Statistics Canada

Double Sampling

M.A. HIDIROGLOU[1]

ABSTRACT

The theory of double sampling is usually presented under the assumption that one of the samples is nested within the other. This type of sampling is called two-phase sampling. The first-phase sample provides auxiliary information (x) that is relatively inexpensive to obtain, whereas the second-phase sample contains the variables of interest. The first-phase data are used in various ways: (a) to stratify the second-phase sample; (b) to improve the estimate using a difference, ratio or regression estimator; or (c) to draw a sub-sample of non-respondent units. However, it is not necessary for one of the samples to be nested in the other or selected from the same frame. The case of *non-nested* double sampling is dealt with in passing in the classical works on sampling (Des Raj 1968, Cochran 1977). This method is now used in several national statistical agencies.

This paper consolidates double sampling by presenting it in a unified manner. Several examples of surveys used at Statistics Canada illustrate this unification.

KEY WORDS : Double sampling ; Auxiliary data ; Regression ; Optimal.

1. INTRODUCTION

The theory of double-phase sampling is usually presented under the assumption that one of the samples is nested within the other. This type of sampling is called two-phase sampling. The first-phase sample provides auxiliary information (x) that is relatively inexpensive to obtain, whereas the second-phase sample contains the variables of interest. The first-phase data are used in various ways: (a) to stratify the second-phase sample; (b) to improve the estimation by using a difference, ratio or regression estimator; or (c) to draw a sub-sample of non-respondent units. Two-phase sampling is a powerful and cost-effective technique with a long history. Neyman (1938) was first to propose it. Rao (1973) studied double sampling in the context of stratification and analytic studies. Cochran (1977) presented the basic results of two-phase sampling, including the simplest regression estimators for this type of sampling design. More recent work on the subject includes that of Breidt and Fuller (1993), who developed efficient estimation methods for three-phase sampling computations using auxiliary data. Chaudhuri and Roy (1994) focused on the optimal properties of simpler but well-known regression estimators of two-phase sampling. Hidiroglou and Särndal (1998) proposed estimators based on calibration and regression for two-phase sampling to account for the availability of auxiliary data at both levels of the sampling design.

Estimation for nested and non-nested double sampling has been treated separately in the survey literature. However, it is not necessary for one of the samples to be nested within the other, or even be selected from the same survey frame. This case will be termed *non-nested* double sampling. It has been briefly discussed in such classical books on sampling such as Des Raj (1968) and Cochran (1977). This method is used in several statistical agencies. For example, at Statistics Canada, the Canadian Survey of Employment, Payrolls and Hours (SEPH) is using this sampling procedure (Rancourt and Hidiroglou 1998). In this survey, two independent samples are drawn from two different frames, which nevertheless represent the same universe. The auxiliary data (x), which includes the number of employees and the total amount of payrolls are obtained from a sample selected from a Canada Customs and Revenue Agency administrative data file. These same variables, together with the variables of interest (y), the number of hours worked by employees and summarised earnings, are collected from a sample drawn from the Statistics Canada Business Register. Another example described by Deville (1999) is the case of a household survey conducted at INSEE.

A single estimator can represent the overall estimation process, and the only difference is with respect to variance estimation. This paper is structured as follows. Part 2 sets out the notation. Part 3 describes how the double sampling procedures can be obtained from a single estimator. In Part 4, the estimated variance for the nested and non-nested calibration estimator is presented. Several practical examples are provided in Part 5. Finally, Part 6 contains a brief summary.

2. NOTATION

2.1 Nested Case

The population is represented by $U = \{1, ..., k, ..., N\}$. First, a probability sample s_1 ($s_1 \subset U$) is selected from population U using a sampling design with inclusion

[1] M.A. Hidiroglou, Business Survey Methods Division, R.H. Coats Building, 11th Floor, Section A, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6. E-mail: hidirog@statcan.ca.

probability of $\pi_{1k} = P(k \varepsilon s_1)$ for the k-th sampled unit in s_1. Given s_1, a second sample s_2 ($s_2 \subseteq s_1 \subseteq U$) is drawn from s_1 using a sample design with conditional inclusion probability $\pi_{2k|s_1} = P(k \varepsilon s_2 \mid s_1)$ for the k-th sampled unit in s_2. Note that the probabilities are conditional since it is assumed that s_1 is known. Figure 1 displays an example of nested sampling.

We assume that $\pi_{1k} > 0$ for all values $k \varepsilon U$ and that $\pi_{2k|s_1} > 0$ for all values $k \varepsilon s_1$. The weight of a sampled unit k will be denoted by $w_{1k} = 1/\pi_{1k}$ for the first-phase sample and $w_{2k} = 1/\pi_{2k|s_1}$ for the second phase sample. The overall sampling weight of a selected second-phase unit, $k \varepsilon s_2$, will therefore be $w_k^* = w_{1k} w_{2k}$.

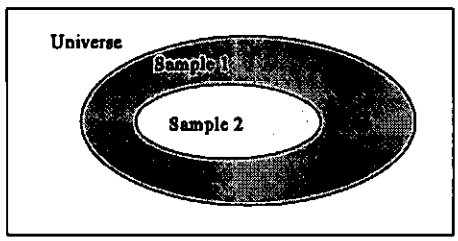

Figure 1. Nested Samples

Let x denote the auxiliary data vector available with the first-phase sample, and x_k the value for unit k. We proceed as in Hidiroglou and Särndal (1998), that is, we divide x_k into two parts x_{1k} and x_{2k}. The values of the data vector x_{1k} as assumed to be known for the entire population U, while the values of data vector x_{2k} are only known for the first-phase sample s_1.

2.2 Non-nested Case

It is possible for the two samples to be drawn independently from the same frame or even from different (but equivalent) frames. Figures 2 and 3 provide examples of these non-nested cases.

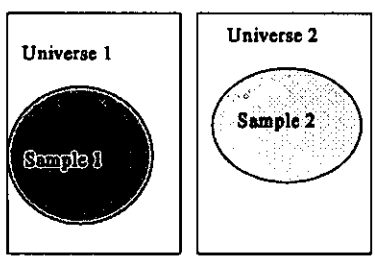

Figure 2. Two independent samples selected from different sample frames

The non-nested case represented by Figure 3 is not considered in this paper. This case can be complicated for arbitrary sampling plans because it is necessary to compute joint inclusion probabilities between the two samples s_1 and s_2. This computation is simpler when the two samples s_1 and s_2 have been selected using a simple sampling design such as simple random sampling (with or without replacement). It is then possible to use Tam's results (1984) to obtain the required joint selection probabilities for the computation of the estimated variance for a given estimator of the total $Y = \Sigma_U y_k$.

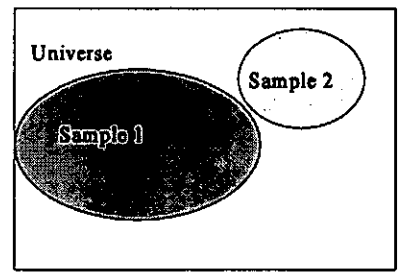

Figure 3. Two samples drawn independently from the same sample frame

For the case that we will study, we assume that samples s_1 and s_2 are drawn independently from two different frames $U_1 = \{1, ..., k, ..., N_1\}$ and $U_2 = \{1, ..., k, ..., N_2\}$ (see Figure 2). The inclusion probabilities of a sampled unit k are respectively $\pi_{1k}^{(1)} = P(k \varepsilon s_1) > 0$ and $\pi_{2k}^{(2)} = P(k \varepsilon s_2) > 0$ for samples s_1 ($s_1 \subseteq U_1$) and s_2 ($s_2 \subseteq U_2$). The weight of unit k is $w_{1k}^{(1)} = 1/\pi_{(1k)}^{(1)}$ for the first sample s_1 and $w_{2k}^{(2)} = 1/\pi_{2k}^{(2)}$ for the second sample s_2. The superscripts (1) and (2) are used to differentiate between the selection probabilities of the samples drawn in the nested case. The sampling units may differ between the two frames, but these frames represent the same coverage. Examples of such sampling procedures were mentioned in the introduction and more details are provided in the second example given in section 5.3.

Let $x_k = (x_{1k}', x_{2k}')'$, be an auxiliary data vector. We assume that $x_{1k}^{(1)}$ is known for all units belonging to frame U_1, while $x_k^{(1)}$ is only known for sample s_1. We collect $y_k^{(2)}, x_k^{(2)}$ from sample s_2. The x data collected for corresponding units in samples s_1 and s_2 may differ. The degree in difference between the data values will vary according to the complexity of the sampling unit, and how much these units differ in concept between the two sampling frames. For « simpler » units the data reported for « similar » units in s_1 and s_2 should be equal or almost equal. Departures in the data similarity for the same units in s_1 and s_2 would most likely be due to the different questionnaire wording or due to different respondents filling in the questionnaires. Nevertheless, we assume that $X_1 = \Sigma_{U_1} x_{1k}^{(1)} = \Sigma_{U_2} x_{1k}^{(2)}$ since U_1 and U_2 have the same coverage.

3. OPTIMAL ESTIMATOR FOR NESTED AND NON-NESTED SAMPLES

In both cases, nested and non-nested, the objective is to estimate the population total $Y = \Sigma_U y_k$ where y_k represents the value of unit $k \varepsilon U$. An unbiased estimator of Y is $\hat{Y}_{HT} = \Sigma_{s_2} w_k^* y_k$, where $w_k^* = w_{1k} w_{2k}$ for the nested case and $w_k^* = w_{2k}^{(2)}$ for the non-nested case.

The sampling weight of a unit is modified by multiplying it by the calibration factor obtained using the various levels of the auxiliary data (universe, first-phase sample). The product is called a "calibration weight". Table 1 summarises the available data for the nested and non-nested cases, corresponding to Figures 1 and 2.

Table 1
Data Available for the Population and Samples

Set of Elements	Nested Case	Non-nested Case
Population	x_{1k}: known for $k \in U$	$x_{1k}^{(1)}$: known for $k \in U_1$
First sample	x_k: observed for $k \in s_1$	$x_k^{(1)}$: observed for $k \in s_1$
Second sample	y_k, x_k: observed for $k \in s_2$	$y_k^{(2)}, x_k^{(2)}$: observed for $k \in s_2$

The following regression estimator is used to estimate the population total Y for nested and non-nested samples:

$$\hat{\hat{Y}}_{REG} = \hat{\hat{Y}}_{HT} + (\boldsymbol{X}_1 - \hat{\boldsymbol{X}}_1)' \boldsymbol{B}_1 + (\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}})' \boldsymbol{B}. \quad (3.1)$$

The various totals corresponding to the auxiliary data x and y-variable of interest given in equation (3.1) are provided in Table 2.

It is assumed that the variances, $V(\hat{\hat{Y}}_{HT})$, and covariances $\mathrm{Cov}(\hat{\boldsymbol{X}}, \hat{\hat{\boldsymbol{X}}}')$, $\mathrm{Cov}(\hat{\boldsymbol{X}}_1, \hat{\boldsymbol{X}}')$, $\mathrm{Cov}(\hat{\boldsymbol{X}}_1, \hat{\hat{\boldsymbol{X}}}')$, $\mathrm{Cov}(\hat{\hat{Y}}_{HT}, \hat{\boldsymbol{X}}')$ and $\mathrm{Cov}(\hat{\hat{Y}}_{HT}, \hat{\hat{\boldsymbol{X}}}')$, are known or estimable.

To simplify the notation, we drop the superscripts for the remainder of this section. The estimation of the parameters, $\boldsymbol{B}$ and $\boldsymbol{B}_1$ as well as of their associated variance, reflect that we have sampled differently for the nested and non-nested cases. The estimators of $\boldsymbol{B}$ and $\boldsymbol{B}_1$ are obtained by minimising the variance of $\hat{\hat{Y}}_{REG}$. This variance is:

$$\begin{aligned} V(\hat{\hat{Y}}_{REG}) = {} & V(\hat{\hat{Y}}_{HT}) + \boldsymbol{B}_1' V(\hat{\boldsymbol{X}}_1)\boldsymbol{B}_1 + \boldsymbol{B}' V(\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}})\boldsymbol{B} \\ & - 2\,\mathrm{Cov}(\hat{\hat{Y}}_{HT}, \hat{\boldsymbol{X}}_1')\boldsymbol{B}_1 + 2\,\mathrm{Cov}(\hat{\hat{Y}}_{HT}, (\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}})')\boldsymbol{B} \\ & - 2\,\boldsymbol{B}_1' \mathrm{Cov}(\hat{\boldsymbol{X}}_1, (\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}})')\boldsymbol{B}. \end{aligned} \quad (3.2)$$

Deriving (3.2) with respect to $\boldsymbol{B}$ and $\boldsymbol{B}_1$, we obtain the following two equations:

$$\begin{aligned} & V(\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}})\boldsymbol{B} + \mathrm{Cov}((\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}}), \hat{\hat{Y}}_{HT}) \\ & \quad - \mathrm{Cov}((\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}}), \hat{\boldsymbol{X}}_1')\boldsymbol{B}_1 = \boldsymbol{0} \end{aligned} \quad (3.3)$$

and

$$-\mathrm{Cov}(\hat{\boldsymbol{X}}_1, (\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}})')\boldsymbol{B} - \mathrm{Cov}(\hat{\boldsymbol{X}}_1, \hat{\hat{Y}}_{HT}) + V(\hat{\boldsymbol{X}}_1)\boldsymbol{B}_1 = \boldsymbol{0}. \quad (3.4)$$

Solving the system of equations (3.3) and (3.4), we obtain the required parameters $\boldsymbol{B}$ **and** $\boldsymbol{B}_1$. That is:

$$\boldsymbol{B} = \boldsymbol{T}^{-1}\boldsymbol{H} \quad (3.5)$$

where

$$\begin{aligned} \boldsymbol{T} = {} & V(\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}}) - (\mathrm{Cov}(\hat{\boldsymbol{X}}_1, (\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}})'))' \\ & V^{-1}(\hat{\boldsymbol{X}}_1)(\mathrm{Cov}(\hat{\boldsymbol{X}}_1, (\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}})')), \end{aligned}$$

$$\begin{aligned} \boldsymbol{H} = {} & (\mathrm{Cov}((\hat{\hat{\boldsymbol{X}}} - \hat{\boldsymbol{X}}), \hat{\hat{Y}}_{HT})) \\ & + (\mathrm{Cov}(\hat{\boldsymbol{X}}_1, (\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}})'))' V^{-1}(\hat{\boldsymbol{X}}_1)\mathrm{Cov}(\hat{\boldsymbol{X}}_1, \hat{\hat{Y}}_{HT}) \end{aligned}$$

and

$$\boldsymbol{B}_1 = \boldsymbol{T}_1^{-1}\boldsymbol{H}_1 \quad (3.6)$$

where

$$\boldsymbol{T}_1 = V(\hat{\boldsymbol{X}}_1),$$

and

$$\boldsymbol{H}_1 = \mathrm{Cov}(\hat{\boldsymbol{X}}_1, \hat{\hat{Y}}_{HT}) + \mathrm{Cov}(\hat{\boldsymbol{X}}_1, (\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}})')'\boldsymbol{B}.$$

Table 2
Sums of the Auxiliary Data x and y for Nested and Non-nested Cases

Set of Elements	Nested Case	Non-nested Case
Population	$X_1 = \sum_U x_{1k}$	$X_1 = \sum_{U_1} x_{1k}^{(1)}$
First sample	$\hat{X}_1 = \sum_{s_1} w_{1k} x_{1k}$; $\hat{X} = \sum_{s_1} w_{1k} x_k$	$\hat{X}_1 = \sum_{s_1} w_{1k} x_{1k}^{(1)}$; $\hat{X} = \sum_{s_1} w_{1k} x_k^{(1)}$
Second sample	$\hat{\hat{X}}_1 = \sum_{s_2} w_k^* x_{1k}$; $\hat{\hat{X}} = \sum_{s_2} w_k^* x_k$ $\hat{\hat{Y}}_{HT} = \sum_{s_2} w_k^* y_k$	$\hat{\hat{X}}_1 = \sum_{s_2} w_{2k} x_{1k}^{(2)}$; $\hat{\hat{X}} = \sum_{s_2} w_{2k} x_k^{(2)}$ $\hat{\hat{Y}}_{HT} = \sum_{s_2} w_{2k}^* y_k^{(2)}$

Result 1: An optimal regression estimator for the nested and non-nested samples is:

$$\hat{\bar{Y}}_{OPT} = \hat{\bar{Y}}_{HT} + (\bar{X}_1 - \hat{\bar{X}}_1)' \hat{B}_{1,OPT} + (\hat{\bar{X}} - \hat{\bar{X}})' \hat{\bar{B}}_{OPT} \qquad (3.7)$$

where

$$\hat{\bar{B}}_{OPT} = \hat{T}^{-1} \hat{H} \qquad (3.8)$$

and

$$\hat{\bar{B}}_{1,OPT} = \hat{T}_1^{-1} \hat{H}_1. \qquad (3.9)$$

$\hat{T}_1, \hat{H}_1, \hat{T}$ and $\hat{H}$ are the estimated values of T_1, H_1, T and H, and they are obtained using a framework leading to the inference based on the sampling design. These values are dependent on the sample selection scheme. The population variance of $\hat{\bar{Y}}_{OPT}$ and its associated estimated variance depend on whether or not the samples are nested or non-nested. Since the regression vectors are optimal, it follows that the regression estimator $\hat{\bar{Y}}_{OPT}$ is also optimal. The optimal form has been discussed by Montanari (1987, 1998, and 2000) for the case of a single phase sampling design.

3.1 The Case of Nested Double Sampling

The theory for this case is developed using a conditional approach. Suppose that two parameters are given by θ_1 and θ_2, and that they are estimated by $\hat{\bar{\theta}}_1$ and $\hat{\bar{\theta}}_2$ from sample s_2. If we condition on the realised sample s_1, then the following well-known results hold:

(i) The expectation of $\hat{\bar{\theta}}$ is $E(\hat{\bar{\theta}}) = E_1 E_2(\hat{\bar{\theta}} \mid s_1)$, where E_2 denotes the expectation of $\hat{\bar{\theta}}$ given s_1.

(ii) The variance of $\hat{\bar{\theta}}$ is

$$V(\hat{\bar{\theta}}) = E_1 V_2(\hat{\bar{\theta}} \mid s_1) + V_1 E_2(\hat{\bar{\theta}} \mid s_1). \qquad (3.10)$$

(iii) The covariance between $\hat{\bar{\theta}}_1$ and $\hat{\bar{\theta}}_2$ is:

$$\mathrm{Cov}(\hat{\bar{\theta}}_1, \hat{\bar{\theta}}_2') = E_1 \mathrm{Cov}_2\left((\hat{\bar{\theta}}_1, \hat{\bar{\theta}}_2') \mid s_1\right) + \mathrm{Cov}_1\left(E_2(\hat{\bar{\theta}}_1 \mid s_1), E_2(\hat{\bar{\theta}}_2' \mid s_1)\right).$$

The various components of $\hat{T}, \hat{H}, \hat{T}_1$ and of $\hat{H}_1$ will be estimated assuming an arbitrary sampling design with a non-fixed sample size. The case of a fixed size sampling design follows easily as it is a special case of the arbitrary sampling design. Using expressions (i) – (iii), we can re-express the terms defining parameter B as:

$$\mathrm{Cov}(\hat{\bar{X}}, \hat{\bar{X}}') = \mathrm{Cov}(\hat{\bar{X}}, \hat{\bar{X}}') = V(\hat{\bar{X}});$$

$$\mathrm{Cov}(\hat{\bar{Y}}_{HT}, \hat{\bar{X}}') = \mathrm{Cov}(\hat{\bar{Y}}_{HT}, \hat{\bar{X}}');$$

$$V(\hat{\bar{X}} - \hat{\bar{X}}) = E_1\left[\sum\sum_{s_1} c_{2k\ell \mid s_1} x_k x_\ell'\right];$$

$$\mathrm{Cov}\left[\hat{\bar{X}}_1, (\hat{\bar{X}} - \hat{\bar{X}})'\right] = \mathbf{0};$$

and

$$\mathrm{Cov}(\hat{\bar{X}}, \hat{\bar{Y}}_{HT}) = \mathrm{Cov}(\hat{\bar{X}}, \hat{\bar{Y}}_{HT}) + E_1\left[\sum\sum_{s_1} c_{2k\ell \mid s_1} x_k x_\ell'\right]; \qquad (3.11)$$

where $c_{2k\ell \mid s_1} = (\pi_{2k\ell \mid s_1} - \pi_{2k \mid s_1}\pi_{2\ell \mid s_1})/\pi_k^* \pi_\ell^*$ and $\hat{\bar{Y}}_{HT} = \sum_{s_1} y_k / \pi_{1k}$. The inclusion probabilities in these expressions are $\pi_{2k\ell \mid s_1} = \Pr(k, \ell \in s_2 \mid s_1)$ and $\pi_k^* = \pi_{1k}\pi_{2k \mid s_1}$.
We can express B more simply as:

$$B = \left[E_1\left(\sum\sum_{s_1} c_{2k\ell \mid s_1} x_k x_\ell'\right)\right]^{-1} E_1\left[\sum\sum_{s_1} c_{2k\ell \mid s_1} x_k y_\ell\right] \qquad (3.12)$$

and the corresponding optimal estimator is given by:

$$\hat{\bar{B}}_{OPT} = \left[\sum\sum_{s_2} \hat{c}_{2k\ell \mid s_1} x_k x_\ell'\right]^{-1} \left[\sum\sum_{s_2} \hat{c}_{2k\ell \mid s_1} x_k y_\ell\right] \qquad (3.13)$$

where $\hat{c}_{2k\ell \mid s_1} = c_{2k\ell \mid s_1} / \pi_{2k\ell \mid s_1}$.
The optimal regression estimator $\hat{\bar{B}}_{1,OPT}$, is given by (3.9) with

$$\hat{T}_1 = \hat{V}(\hat{\bar{X}}_1)$$

and

$$\hat{H}_1 = \hat{\mathrm{Cov}}(\hat{\bar{X}}_1, \hat{\bar{Y}}_{HT}) + \hat{\mathrm{Cov}}(\hat{\bar{X}}_1, \hat{\bar{X}}')\hat{\bar{B}}_{OPT} - \hat{\mathrm{Cov}}(\hat{\bar{X}}_1, \hat{\bar{X}}')\hat{\bar{B}}_{OPT}.$$

Each component defining $\hat{T}_1$ and $\hat{H}_1$ is estimated as follows. We first estimate $V(\hat{\bar{X}}_1) = \sum\sum_{s_1} c_{1k\ell} x_{1k} x_{1\ell}'$ by

$$\hat{V}(\hat{\bar{X}}_1) = \sum\sum_{s_1} \hat{c}_{1k\ell} x_{1k} x_{1\ell}' \qquad (3.14)$$

where $c_{1k\ell} = (\pi_{1k\ell} - \pi_{1k}\pi_{1\ell})/(\pi_{1k}\pi_{1\ell})$ and $\hat{c}_{1k\ell} = c_{1k\ell}/\pi_{1k\ell}$.

Next, since

$$\begin{aligned}\mathrm{Cov}\left(\hat{X}_1, \hat{\hat{Y}}_{\mathrm{HT}}\right) &= E_1\,\mathrm{Cov}_2\left[\left(\hat{X}_1, \hat{\hat{Y}}_{\mathrm{HT}}\right)\middle| s_1\right] \\ &\quad + \mathrm{Cov}_1\left[E_2\left(\hat{X}_1 \middle| s_1\right), E_2\left(\hat{\hat{Y}}_{\mathrm{HT}} \middle| s_1\right)\right] \\ &= \mathrm{Cov}_1\left(\hat{X}_1, \hat{Y}_{\mathrm{HT}}\right) \\ &= \sum \sum_{U_1} c_{1k\ell} x_{1k} y_{1\ell} \qquad (3.15)\end{aligned}$$

we estimate $\mathrm{Cov}\,(\hat{X}_1, \hat{\hat{Y}}_{\mathrm{HT}})$ by

$$\hat{\mathrm{Cov}}\left(\hat{X}_1, \hat{\hat{Y}}_{\mathrm{HT}}\right) = \sum \sum_{s_2} c^*_{1k\ell} x_{1k} y_\ell \qquad (3.16)$$

where

$$c^*_{1k\ell} = c_{1k\ell}/\pi^*_{k\ell},\ \pi^*_{k\ell} = \pi_{1k\ell}\,\pi_{2k\ell|s_1},$$
$$\pi_{1k\ell} = \Pr\,(k, \ell\,\varepsilon\, s_1),$$
$$\pi_{2k\ell|s_1} = \Pr\,(k, \ell\,\varepsilon\, s_2 \,|\, s_1)$$
$$\text{and}\quad \pi^*_k = \pi_{1k}\,\pi_{2k|s_1}.$$

Similarly,

$$\hat{\mathrm{Cov}}\left(\hat{X}_1, \hat{\hat{X}}'\right) = \sum \sum_{s_2} c^*_{1k\ell} x_{1k} y'_\ell \qquad (3.17)$$

and

$$\hat{\mathrm{Cov}}\left(\hat{X}_1, \hat{X}'\right) = \sum \sum_{s_1} \hat{c}_{1k\ell} x_{1k} x'_\ell. \qquad (3.18)$$

Hence, in the case of nested double sampling the optimal estimator of $\boldsymbol{B}_1$ is given by:

$$\begin{aligned}\hat{\boldsymbol{B}}_{1,\mathrm{OPT}} &= \left(\hat{V}\left(\hat{X}_1\right)\right)^{-1}\left[\hat{\mathrm{Cov}}\left(\hat{X}_1, \hat{\hat{Y}}_{\mathrm{HT}}\right)\right. \\ &\quad \left. + \left(\hat{\mathrm{Cov}}\left(\hat{X}_1, \hat{X}'\right) - \hat{\mathrm{Cov}}\left(\hat{X}_1, \hat{\hat{X}}\right)\right)\hat{\hat{\boldsymbol{B}}}_{\mathrm{OPT}}\right] \qquad (3.19)\end{aligned}$$

where the components of $\hat{\boldsymbol{B}}_{1,\mathrm{OPT}}$ have been defined by expressions (3.14) – (3.18).

The optimal form of estimators $\hat{\boldsymbol{B}}_{1,\mathrm{OPT}}$ and $\hat{\hat{\boldsymbol{B}}}_{\mathrm{OPT}}$ has its advantages and disadvantages. One of the biggest advantages of the optimal form, as reported by Cassady and Valliant (1993), Rao (1994), and Montanari (2000), is that it has good conditional inference properties (by conditioning on the auxiliary variable $\boldsymbol{x}$). As Montanari (2000) observed, the asymptotic optimality of $\hat{\hat{Y}}_{\mathrm{OPT}}$ is strictly a property based on the sampling design and achieved conditionally on the finite population. The biggest disadvantage of the optimal estimator is that it requires the computation of joint inclusion probabilities.

We can, however, use the optimal form, and express it more simply for several sampling designs. For sampling designs where the sample selection is with unequal probability and without replacement, we can bypass the computation of the joint probability by approximating the exact variance. Several authors, including Hartley and Rao (1962), Deville (1999), Berger (1998), Rósen (2000) and Brewer (2000) proposed such approximating procedures. Recently, Tillé (2001) proposed the following approximation for the estimated variance of $\hat{Y}_{\mathrm{HT}} = \sum_s y_k/\pi_k$ in the context of single-phase sampling, where

$$\begin{aligned}\hat{V}\left(\hat{Y}_{\mathrm{HT}}\right) &= \sum_s \frac{c_k}{\pi_k^2}\left(y_k - y_k^*\right)^2 \\ &= \sum_s c_k\left(\frac{y_k}{\pi_k} - \tilde{y}\right)^2. \qquad (3.20)\end{aligned}$$

Here, c_k is the variable used as the approximation, $y_k^* = \pi_k \sum_s c_\ell\, y_\ell/\pi_\ell/\sum_s c_\ell$, $\tilde{y} = y_k^*/\pi_k$, and π_k is the probability of selection of a given unit k. Tillé (2001) provided several examples of the c_k values for various sampling schemes.

This formula is exact in the case of a stratified simple sampling design drawn without replacement in each stratum $U_h\,(h = 1, ..., L)$ of population U. Let k be a sampled unit in sample s_h from stratum U_h, then $c_k = n_h/(n_h - 1)(1 - n_h/N_h)$ if $k\,\varepsilon\, U_h$ and 0 otherwise, and $\pi_k = n_h/N_h$ if $k\,\varepsilon\, U_h$ and 0 otherwise. This gives us the exact estimated variance, $\hat{V} = (\hat{Y}_{\mathrm{HT}}) = \sum_{h=1}^{L} N_h^2(1 - n_h/N_h)\,\sum_{s_h}(y_k - \bar{y}_h)^2/n_h(n_h - 1)$. The formula is also exact in the case of a stratified sampling design where the sample is selected with replacement. Here $c_k = 1$ for all units belonging to stratum U_h and zero otherwise. Using this approximation, the double sums appearing in $\hat{\hat{\boldsymbol{B}}}_{\mathrm{OPT}}$ and $\hat{\boldsymbol{B}}_{1,\mathrm{OPT}}$ can be expressed as simple sums. Hidiroglou and Särndal (1998) bypassed the problem of double sums in estimating $\boldsymbol{B}$ and $\boldsymbol{B}_1$ by proposing the GREG estimator, $\hat{\hat{Y}}_{\mathrm{GREG}}$, for a nested two-phase sampling design. Their estimator is given by:

$$\hat{\hat{Y}}_{\mathrm{GREG}} = \hat{\hat{Y}}_{\mathrm{HT}} + \left(\boldsymbol{X}_1 - \hat{\boldsymbol{X}}_1\right)'\hat{\boldsymbol{B}}_{1,\mathrm{GREG}} + \left(\hat{\boldsymbol{X}} - \hat{\hat{\boldsymbol{X}}}\right)'\hat{\hat{\boldsymbol{B}}}_{\mathrm{GREG}}$$

where

$$\hat{\hat{\boldsymbol{B}}}_{\mathrm{GREG}} = \left(\sum_{s_2} \frac{\tilde{w}_{1k} w_{2k} \boldsymbol{x}_k \boldsymbol{x}_k'}{\sigma_{2k}^2}\right)^{-1} \sum_{s_2} \frac{\tilde{w}_{1k} w_{2k} \boldsymbol{x}_k y_k}{\sigma_{2k}^2}, \qquad (3.21)$$

$$\begin{aligned}\hat{\boldsymbol{B}}_{1,\mathrm{GREG}} &= \left(\sum_{s_1} \frac{w_{1k} \boldsymbol{x}_{1k} \boldsymbol{x}_{1k}'}{\sigma_{1k}^2}\right)^{-1} \\ &\quad \left\{\sum_{s_2} \frac{w_k^* \boldsymbol{x}_{1k} y_k}{\sigma_{1k}^2} + \sum_{s_1} \frac{w_{1k} \boldsymbol{x}_{1k} \boldsymbol{x}_k'}{\sigma_{1k}^2}\hat{\hat{\boldsymbol{B}}}_{\mathrm{GREG}}\right. \\ &\quad \left. - \sum_{s_2} \frac{w_k^* \boldsymbol{x}_{1k} \boldsymbol{x}_k'}{\sigma_{1k}^2}\hat{\hat{\boldsymbol{B}}}_{\mathrm{GREG}}\right\} \qquad (3.22)\end{aligned}$$

with $\{\sigma_{1k}^2 : k\,\varepsilon\, s_1\}$ and $\{\sigma_{2k}^2 : k\,\varepsilon\, s_2\}$ being predetermined positive factors.

Estimators $\hat{B}_{\text{GREG}}$ and $\hat{B}_{1,\text{GREG}}$ can be justified either by assuming different regression models for each phase or by using two successive calibrations. For the calibration approach, calibration weights $\tilde{w}_{1k}$ associated with the first-phase are first obtained, and they satisfy the calibration equation $\Sigma_{s_1} \tilde{w}_{1k} x_{1k} = \Sigma_U x_{1k}$. These calibration weights can be expressed as the product of sample weights w_{1k} and a calibration factor g_{1k} where:

$$g_{1k} = 1 + \left(\sum_U x_{1k} - \sum_{s_1} w_{1k} x_{1k}\right)' \left(\sum_{s_1} w_{1k} \frac{x_{1k} x_{1k}'}{\sigma_{1k}^2}\right)^{-1} \frac{x_{1k}}{\sigma_{1k}^2} \qquad (3.23)$$

for $k \varepsilon s_1$.

The first-phase calibration weights $\tilde{w}_{1k}$ are then used as initial weights to compute the overall calibration weights $\tilde{w}_k^*$. These overall calibration weights satisfy the second-phase calibration equation $\Sigma_{s_2} \tilde{w}_{1k}^* x_k = \Sigma_{s_1} \tilde{w}_{1k} x_k$. The estimator of the total, $\hat{Y}_{\text{GREG}}$, can be expressed as the sum of the product of the overall calibration weight $\tilde{w}_k^*$ and the associated y-value, that is $\hat{Y}_{\text{GREG}} = \Sigma_{s_2} \tilde{w}_k^* y_k$. The calibrated overall weights can be expressed as $\tilde{w}_k^* = w_k^* g_k^*$, where $g_k^* = g_{1k} g_{2k}$. Here, g_{1k} is given by (3.23), while g_{2k} is equal to

$$g_{2k} = 1 + \left(\sum_{s_1} \tilde{w}_{1k} x_k - \sum_{s_2} \tilde{w}_{1k} w_{2k} x_k\right)' \left(\sum_{s_1} \frac{\tilde{w}_{1k} w_{2k} x_k x_k'}{\sigma_{2k}^2}\right)^{-1} \frac{x_k}{\sigma_{2k}^2} \qquad (3.24)$$

for $k \varepsilon s_2$.

Comment: The estimators of $\hat{B}_{1,\text{GREG}}$ (3.21) and $\hat{\tilde{B}}_{\text{GREG}}$ (3.22) correspond to Hidiroglou and Särndal's (1998) *additive case* and have the same form as the optimal regression estimators $\hat{B}_{1,\text{OPT}}$ (3.8) and $\hat{\tilde{B}}_{\text{OPT}}$ (3.9). Indeed, the components of the estimator of B are obtained by respectively estimating T by $(\Sigma_{s_2} w_{1k} w_{2k} x_k x_k'/\sigma_{2k}^2)$ and H by $\Sigma_{s_2} w_{1k} w_{2k} x_k y_k/\sigma_{2k}^2$. The second terms of H and T are exactly equal to zero. Similarly, to estimate B_1, the component T_1 is estimated by $\Sigma_{s_1} w_{1k} x_{1k} x_{1k}'/\sigma_{1k}^2$, while H_1 is estimated by

$$\sum_{s_2} \frac{w_k^* x_{1k} y_{2k}}{\sigma_{1k}^2} + \left(\sum_{s_1} \frac{w_{1k} x_{1k} x_k'}{\sigma_{1k}^2} - \sum_{s_2} \frac{w_k^* x_{1k} x_k'}{\sigma_{1k}^2}\right)' \hat{\tilde{B}}_{\text{GREG}}.$$

The estimated variance of $\hat{Y}_{\text{GREG}} = \hat{Y}_{\text{HT}} + (X_1 - \hat{X}_1)' \hat{B}_{1,\text{GREG}} + (\hat{X} - \hat{\hat{X}})' \hat{\tilde{B}}_{\text{GREG}}$ is presented in Hidiroglou and Särndal (1998).

Comment: The efficiency of the GREG, as stated in Särndal, Swensson and Wretman (1992), requires that the proposed model be correct. Furthermore, if the sample size is large enough, optimal estimators are more efficient (Rao 1994) than the GREG. However, if the sample size is relatively small, one disadvantage of the optimal form OPT is that it is generally less stable and more complex to compute than the GREG. Furthermore, an additional consequence of a relatively small sample size, as reported by Särndal (1996), and illustrated by simulation by Montanari (2000), is that if the sample size is relatively small, then the optimal form is not significantly more efficient than the GREG. It is even possible for the estimated variance to be greater than that associated with the GREG.

3.2 The Case of Non-nested Double Sampling

Deville (1999) considered the non-nested case (Figure 2) by assuming that x_{2k} is known for s_1 and s_2. The optimal regression estimator is:

$$\hat{Y}_{\text{OPT}} = \hat{Y}_{\text{HT}} + \left(\hat{X}_2 - \hat{\hat{X}}_2\right)' \hat{\tilde{B}}_{2,\text{OPT}} \qquad (3.25)$$

where $\hat{Y}_{\text{HT}} = \Sigma_{s_2} w_{2k} y_k^{(2)}$, $\hat{X}_2 = \Sigma_{s_1} w_{1k} x_{2k}^{(1)}$, $\hat{\hat{X}}_2 = \Sigma_{s_2} w_{2k} x_{2k}^{(2)}$. The optimal estimator for $B_2 = (\Sigma_{U_2} x_{2k} x_{2k}')^{-1} \Sigma_{U_2} x_{2k} y_k$ is $\hat{\tilde{B}}_{2,\text{OPT}} = (\hat{V}(\hat{X}_2) + \hat{V}(\hat{\hat{X}}_2))^{-1}\, \hat{\text{Cov}}(\hat{Y}_{\text{HT}}, \hat{\hat{X}}_2')$ if the two sampling frames U_1 and U_2 are independent. The form of the variance and of the covariance terms defining $\hat{\tilde{B}}_{2,\text{OPT}}$ depends on the sampling design of s_1 and s_2.

The accuracy of the estimator of X_2 can be improved by minimising the variance of $\tilde{X}_2 = A_2 \hat{X}_2 + (I - A_2) \hat{\hat{X}}_2$ yielding, $A_2 = (V(\hat{X}_2) + V(\hat{\hat{X}}_2))^{-1} V(\hat{\hat{X}}_2)$. Assuming that $V(\hat{X}_2)$ is approximately a multiple of $V(\hat{\hat{X}}_2)$, that is $V(\hat{X}_2) \doteq \alpha_2 V(\hat{\hat{X}}_2)$, we obtain $A_2 \doteq I/(1 + \alpha_2)$ where I is the identity matrix has the same dimension as the covariance matrix $V(\hat{X}_2)$. The optimal value of α_2 is obtained by minimising the variance of $\tilde{X}_2$. A sub-optimal but adequate choice, suggested by Deville (1999), for α_2 is $\alpha_2 = n_1/(n_1 + n_2)$, where n_1 and n_2 are the respective sizes of samples s_1 and s_2. Note that Korn and Graubart (1999) also made the same suggestion in the context of combining two totals estimated from two different sources. Substituting $\tilde{X}_2$ in place of $\hat{X}_2$ in expression (3.25), yields

$$\tilde{X}_2 - \hat{\hat{X}}_2 = \left(\hat{X}_2 - \hat{\hat{X}}_2\right)/(1 + \alpha_2). \qquad (3.26)$$

The estimator of the population total Y, is:

$$\tilde{\tilde{Y}}_{\text{OPT}} = \hat{Y}_{\text{HT}} + \left(\tilde{X}_2 - \hat{\hat{X}}_2\right)' \tilde{\tilde{B}}_{2,\text{OPT}} \qquad (3.27)$$

where

$$\tilde{\tilde{B}}_{2,\text{OPT}} = -\left[\hat{V}\left(\tilde{X}_2 - \hat{\hat{X}}_2\right)\right]^{-1} \hat{\text{Cov}}\left(\hat{Y}_{\text{HT}}, \left(\tilde{X}_2 - \hat{\hat{X}}_2\right)'\right). \qquad (3.28)$$

If (3.26) is substituted in (3.28), we can re-express $\tilde{\tilde{B}}_{2,\text{OPT}}$ as:

$$\tilde{\tilde{B}}_{2,\text{OPT}} = \left[\hat{V}\left(\hat{\hat{X}}_2\right)\right]^{-1} \hat{\text{Cov}}\left(\hat{Y}_{\text{HT}}, \hat{\hat{X}}_2\right). \qquad (3.29)$$

Comment: We see that $\hat{\bar{Y}}_{OPT}$ (3.25) is exactly equal to $\tilde{\bar{Y}}_{OPT}$ (3.27). This implies that there was no advantage in using a better estimator of X_2 to estimate Y. However, the estimator $\tilde{\tilde{B}}_{2,OPT}$ associated with $\tilde{\tilde{Y}}_{OPT}$ looks more like a traditional regression estimator than the regression estimator $\hat{\tilde{B}}_{2,OPT}$ associated with $\hat{\bar{Y}}_{OPT}$.

Note that the GREG estimator for the case where $\tilde{X}_2$ is used instead of $\hat{X}_2$ is:

$$\tilde{\tilde{Y}}_{GREG} = \hat{\bar{Y}}_{HT} + \left(\tilde{X}_2 - \hat{X}_2\right)' \tilde{\tilde{B}}_{2,GREG} \qquad (3.30)$$

where

$$\tilde{\tilde{B}}_{2,GREG} = \left(\sum_{s_2} w_{2k} x_k^{(2)} x_k^{\prime(2)} / \sigma_{2k}^2\right)^{-1} \sum_{s_2} w_{2k} x_k^{(2)} y_k^{(2)} / \sigma_{2k}^2.$$

Furthermore, if we also know $x_{1k}^{(1)}$ for $k \varepsilon U_1$ where $X_1 = \sum_{U_1} x_{1k}^{(1)}$, we can consider the regression estimator

$$\tilde{\tilde{Y}}_{OPT} = \hat{\bar{Y}}_{HT} + \left(X_1 - \tilde{X}_1\right)' \tilde{B}_{1,OPT} + \left(\tilde{X} - \hat{X}\right)' \tilde{\tilde{B}}_{OPT}. \qquad (3.31)$$

We obtain $\tilde{X}$ by minimising the linear combination $A\hat{X} + (I - A)\hat{\tilde{X}}$ and $V(\hat{X}) \doteq \alpha V(\hat{\tilde{X}})$. The difference between $\tilde{X}$ and $\hat{\tilde{X}}$ can be re-expressed as

$$\tilde{X} - \hat{\tilde{X}} = (\tilde{X} - \hat{X})/(1 + \alpha). \qquad (3.32)$$

Given that s_1 and s_2 are independent samples, it can be shown that:

$$\tilde{\tilde{B}}_{OPT} = \left[\hat{V}(\hat{\tilde{X}})\right]^{-1} \text{ Côv } \left(\hat{\tilde{X}}, \hat{\bar{Y}}_{HT}\right) \qquad (3.33)$$

and that

$$\tilde{B}_{1,OPT} = \left[\hat{V}\left(\hat{\tilde{X}}_1\right)\right]^{-1} \left[\text{Côv } \left(\hat{\tilde{X}}_1, \hat{\bar{Y}}_{HT}\right)\right]. \qquad (3.34)$$

The components of $\tilde{B}_{OPT}$ are estimated by:

$$\hat{V}(\hat{\tilde{X}}) = \sum \sum_{s_2} \hat{c}_{2k\ell} x_k^{(2)} x_\ell^{\prime(2)} \qquad (3.35)$$

and

$$\text{Côv}\left(\hat{\tilde{X}}, \hat{\bar{Y}}_{HT}\right) = \sum \sum_{s_2} \hat{c}_{2k\ell} x_k^{(2)} y_\ell^{(2)} \qquad (3.36)$$

whereas the components of $\tilde{\tilde{B}}_{1,OPT}$ are estimated by:

$$\hat{V}\left(\hat{\tilde{X}}_1\right) = \sum \sum_{s_2} \hat{c}_{2k\ell} x_{1k}^{(2)} x_{1\ell}^{\prime(2)} \qquad (3.37)$$

and

$$\text{Côv}\left(\hat{\tilde{X}}_1, \hat{\bar{Y}}_{HT}\right) = \sum \sum_{s_2} \hat{c}_{2k\ell} x_{1k}^{(2)} y_\ell^{(2)} \qquad (3.38)$$

where

$$\hat{c}_{2k\ell} = \frac{\pi_{2k\ell} - \pi_{2k}\pi_{2\ell}}{(\pi_{2k\ell})(\pi_{2k}\pi_{2\ell})}.$$

Approximation (3.20) can also be used to estimate the terms (3.35) – (3.38). The corresponding GREG which bypasses the computation of joint selection probabilities is given by:

$$\tilde{\tilde{Y}}_{GREG} = \hat{\bar{Y}}_{HT} + \left(X_1 - \tilde{X}_1\right)' \tilde{B}_{1,GREG} + \left(\tilde{X} - \hat{X}\right)' \tilde{\tilde{B}}_{GREG} \qquad (3.39)$$

where $X_1 = \sum_{U_1} x_{1k}^{(1)}$, $\hat{X}_1 = \sum_{s_1} w_{1k} x_{1k}^{(1)}$, $\hat{X} = \sum_{s_1} w_{1k} x_k^{(1)}$ and $\hat{\tilde{X}} = \sum_{s_2} w_{2k} x_k^{(2)}$.

GREG-type regression estimators in equation (3.39) are estimated by

$$\tilde{B}_{1,GREG} = \left(\sum_{s_2} w_{2k} \frac{x_{1k}^{(2)} x_{1k}^{\prime(2)}}{\sigma_{1k}^2}\right)^{-1} \sum_{s_2} w_{2k} \frac{x_{1k}^{(2)} y_k^{(2)}}{\sigma_{1k}^2} \qquad (3.40)$$

and

$$\tilde{\tilde{B}}_{GREG} = \left(\sum_{s_2} w_{2k} \frac{x_k^{(2)} x_k^{\prime(2)}}{\sigma_{2k}^2}\right)^{-1} \sum_{s_2} w_{2k} \frac{x_k^{(2)} y_k^{(2)}}{\sigma_{2k}^2}. \qquad (3.41)$$

4. ESTIMATOR OF THE VARIANCE FOR THE OPTIMAL REGRESSION ESTIMATOR

4.1 Nested Double Sampling

Recall that the optimal regression estimator of Y is given by

$$\hat{\bar{Y}}_{OPT} = \hat{\bar{Y}}_{HT} + \left(X_1 - \hat{X}_1\right)' \hat{B}_{1,OPT} + \left(\hat{X} - \hat{\tilde{X}}\right)' \hat{\tilde{B}}_{OPT}. \qquad (4.1)$$

To obtain the estimated variance of (4.1), we re-express the terms associated with the y-variable within $\hat{\tilde{B}}_{OPT}$ and $\hat{B}_{1,OPT}$ as a simple sums instead of double sums. Montanari (1998) described this algebra for an arbitrary single-phase sampling design. Following Montanari (1998), and adapting the single-phase algebra to double sampling, we obtain:

$$\hat{\tilde{B}}_{OPT} = \left[\sum \sum_{s_2} \hat{c}_{2k\ell|s_1} x_k x_\ell'\right]^{-1} \left[\sum \sum_{s_2} \hat{c}_{2k\ell|s_1} x_k y_\ell\right]$$

$$= \left[\sum \sum_{s_2} \hat{c}_{2k\ell|s_1} x_k x_\ell'\right]^{-1} \left[\sum_{s_2} \frac{a_{2k}}{\pi_k^*} y_k\right] \qquad (4.2)$$

where

$$a_{2k} = \frac{1 - \pi_{2k|s_1}}{\pi_k^*} x_k + \sum_{\substack{\ell \neq k \\ \ell \varepsilon s_2}} \frac{\left(\pi_{2k\ell|s_1} - \pi_{2k|s_1}\pi_{2\ell|s_1}\right)}{\pi_{2k\ell|s_1}\pi_\ell^*} x_\ell.$$

We approximate $\hat{\tilde{B}}_{1,OPT}$ given by (3.15) by $[\hat{V}(\hat{\tilde{X}}_1)]^{-1} [\text{Côv}(\hat{\tilde{X}}_1, \hat{\bar{Y}}_{HT})]$, and hence,

$$\hat{B}_{1,\mathrm{OPT}} \doteq \left[\hat{V}\left(\hat{X}_1\right)\right]^{-1}\left[\hat{\mathrm{Cov}}\left(\hat{X}_1, \hat{Y}_{\mathrm{HT}}\right)\right]$$

$$= \left[\sum\sum_{s_1} \hat{c}_{1k\ell}\, x_{1k} x_{1\ell}'\right]^{-1}\left[\sum_{s_1} \frac{a_{1k}}{\pi_{1k}} y_k\right] \quad (4.3)$$

where

$$a_{1k} = \frac{1-\pi_{1k}}{\pi_{1k}} x_{1k} + \sum_{\substack{\ell \neq k \\ \ell \in s_1}} \frac{\left(\pi_{1k\ell} - \pi_{1k}\pi_{1\ell}\right)}{\pi_{1\ell}\pi_{1k\ell}} x_{1\ell}.$$

By substituting (4.2) and (4.3) in (4.1), and by subtracting the population total Y, we get:

$$\hat{Y}_{\mathrm{OPT}} - Y \doteq \left(\sum_{s_1} g_{1k}\frac{y_k}{\pi_{1k}} - \sum_U y_k\right) + \left(\sum_{s_2} g_{2k}\frac{y_k}{\pi_k^*} - \sum_{s_1}\frac{y_k}{\pi_{1k}}\right) \quad (4.4)$$

where

$$g_{1k} = 1 + \left(X_1 - \hat{X}_1\right)'\left(\hat{V}\left(\hat{X}_1\right)\right)^{-1} a_{1k} \quad \text{for} \quad k \in s_1 \quad (4.5)$$

and

$$g_{2k} = 1 + \left(\hat{X} - \hat{\hat{X}}\right)'\left(\hat{V}\left(\hat{\hat{X}}\right)\right)^{-1} a_{2k} \quad \text{for} \quad k \in s_2. \quad (4.6)$$

Result 2: The estimated variance of $\hat{\hat{Y}}_{\mathrm{OPT}}$ defined by equation (4.1) is:

$$\hat{V}\left(\hat{\hat{Y}}_{\mathrm{OPT}}\right) = \sum\sum_{s_2} c_{1k\ell}^* g_{1k} g_{1\ell} e_{1k} e_{1\ell} + \sum\sum_{s_2} c_{2k\ell}^* g_{2k} g_{2\ell} e_{2k} e_{2\ell} \quad (4.7)$$

where

$$c_{1k\ell}^* = \frac{\left(\pi_{1k\ell} - \pi_{1k}\pi_{1\ell}\right)}{\pi_{k\ell}^* \pi_{1k}\pi_{1\ell}};$$

$$c_{2k\ell}^* = \frac{\left(\pi_{2k\ell|s_1} - \pi_{2k|s_1}\pi_{2\ell|s_1}\right)}{\pi_{2k\ell|s_1}\pi_k^*\pi_\ell^*};$$

$$e_{1k} = y_k - x_{1k}'\hat{B}_{1,\mathrm{OPT}};$$

and

$$e_{2k} = y_k - x_k'\hat{\hat{B}}_{\mathrm{OPT}}.$$

4.2 Non-nested Double Sampling

We obtain the estimated variance of $\tilde{\hat{Y}}_{\mathrm{OPT}}$ by using the following approximation.

$$\tilde{\hat{Y}}_{\mathrm{OPT}} = \hat{\hat{Y}}_{\mathrm{HT}} + \left(X_1 - \tilde{X}_1\right)'\tilde{B}_{1,\mathrm{OPT}} + \left(\tilde{X} - \hat{\hat{X}}\right)'\tilde{\hat{B}}_{\mathrm{OPT}} = \ddot{Y}_{\mathrm{OPT}} + O_p\left(n_1^{-1/2}\right) \quad (4.8)$$

where

$$\ddot{Y}_{\mathrm{OPT}} = \hat{\hat{Y}}_{\mathrm{HT}} + \left(X_1 - \tilde{X}_1\right)' B_{1,\mathrm{OPT}} + \left(\tilde{X} - \hat{\hat{X}}\right)' B_{\mathrm{OPT}}. \quad (4.9)$$

Decomposing $\ddot{Y}_{\mathrm{OPT}}$ into more elementary components, we have that:

$$\ddot{Y}_{\mathrm{OPT}} = \hat{\hat{Y}}_{\mathrm{HT}} + \left(X_1 - \frac{\hat{X}_1 + \alpha\hat{\hat{X}}_1}{1+\alpha}\right)' B_{1,\mathrm{OPT}} + \frac{\left(\tilde{X} - \hat{\hat{X}}\right)'}{1+\alpha} B_{\mathrm{OPT}}$$

$$= \left(\hat{\hat{Y}}_{\mathrm{HT}} - \frac{1}{1+\alpha}\left(\alpha\hat{\hat{X}}_1' B_{1,\mathrm{OPT}} + \hat{\hat{X}}' B_{1,\mathrm{OPT}}\right)\right) + \left(X_1' B_{1,\mathrm{OPT}} - \frac{1}{1+\alpha}\left(\hat{X}_1' B_{1,\mathrm{OPT}} - \hat{\hat{X}}' B_{\mathrm{OPT}}\right)\right). \quad (4.10)$$

The variance of $\ddot{Y}_{\mathrm{OPT}}$ is:

$$V\left(\ddot{Y}_{\mathrm{OPT}}\right) = V\left(\hat{\hat{Y}}_{\mathrm{HT}} - \frac{1}{1+\alpha}\left(\alpha\hat{\hat{X}}_1' B_{1,\mathrm{OPT}} + \hat{\hat{X}}' B_{\mathrm{OPT}}\right)\right) + \frac{1}{(1+\alpha)^2}\left[\alpha B_{1,\mathrm{OPT}}' V\left(\hat{\hat{X}}_1\right) B_{1,\mathrm{OPT}} + B_{\mathrm{OPT}}' V\left(\hat{\hat{X}}\right) B_{\mathrm{OPT}} + 2\alpha\left(B_{\mathrm{OPT}}' V\left(\hat{\hat{X}}\right)\tilde{B}_{1,\mathrm{OPT}}' + \mathrm{Cov}\left(\hat{\hat{X}}_1, \hat{\hat{X}}'\right)\right) B_{\mathrm{OPT}}\right] \quad (4.11)$$

Result 3: The estimated variance of $\tilde{\hat{Y}}_{\mathrm{OPT}}$, $\hat{V}\left(\tilde{\hat{Y}}_{\mathrm{OPT}}\right)$, defined by equation (4.8) is approximately equal to:

$$\hat{V}\left(\hat{\hat{Y}}_{\mathrm{HT}} - \frac{1}{1+\alpha}\left(\alpha\hat{\hat{X}}_1' \tilde{B}_{1,\mathrm{OPT}} + \hat{\hat{X}}'\tilde{\hat{B}}_{\mathrm{OPT}}\right)\right) + \frac{1}{(1+\alpha)^2}\left[\alpha\tilde{B}_{1,\mathrm{OPT}}' \hat{V}\left(\hat{\hat{X}}_1\right)\tilde{B}_{1,\mathrm{OPT}} + \tilde{\hat{B}}_{\mathrm{OPT}}' \hat{V}\left(\hat{\hat{X}}\right)\tilde{\hat{B}}_{\mathrm{OPT}} + 2\alpha\left(\tilde{\hat{B}}_{\mathrm{OPT}}' \hat{V}\left(\hat{\hat{X}}\right)\tilde{B}_{1,\mathrm{OPT}} + \hat{\mathrm{Cov}}\left(\hat{\hat{X}}_1, \hat{\hat{X}}'\right)\tilde{\hat{B}}_{\mathrm{OPT}}\right)\right]. \quad (4.12)$$

Computation of the first term of (4.12) is based on the residuals $y_k - (\alpha x_{1k}'\tilde{B}_{1,\mathrm{OPT}} + x_k'\tilde{\hat{B}}_{\mathrm{OPT}})/(1+\alpha)$. The computation of the other terms of (4.12) is mainly based on the estimated variances of $\hat{X}_1$ and of $\hat{\hat{X}}$, as well as on their estimated covariances. We can use the approximation of the variance, as described by Tillé (2001), and suitably adapt it to estimate the required covariances.

5. SOME SPECIFIC EXAMPLES

Three traditional examples for double sampling are presented for the two cases (nested and non-nested). Furthermore, we briefly describe how two major business surveys carried out by Statistics Canada use double sampling.

5.1 Nested Sampling

Example 1: Let us assume that a simple random sample s_1 of size n_1 is selected from a population U of size N. The sample is stratified into L strata s_{1h} each of size n_{1h}. Random samples s_{2h} of size n_{2h} are then selected without replacement in each stratum s_{1h}. The estimator of the total is $\hat{Y}_{EXP} = N \sum_{h=1}^{L} p_{1h} \bar{y}_{2h} = N \bar{y}_{2,st}$, where $p_{1h} = n_{1h}/n_1$. Using (4.7), we can show that the estimated variance of $\hat{Y}_{EXP}$, $\hat{V}(\hat{Y}_{EXP})$, consists of the sum of $\hat{V}_1(\hat{Y}_{EXP})$ and $\hat{V}_2(\hat{Y}_{EXP})$ corresponding to the first and second phases of the sampling design. Thus:

$$\hat{V}\left(\hat{Y}_{EXP}\right) = \hat{V}_1\left(\hat{Y}_{EXP}\right) + \hat{V}_2\left(\hat{Y}_{EXP}\right)$$

where

$$\hat{V}_1\left(\hat{Y}_{EXP}\right) = N^2 \frac{(1-f_1)}{n_1} \sum_{h=1}^{L} p_{1h}\left[(1-a_h)\hat{S}_{2yh}^2 + \frac{n_1}{n_1-1}\left(\bar{y}_{2h} - \bar{y}_{2,st}\right)^2\right];$$

$$\hat{V}_2\left(\hat{Y}_{EXP}\right) = N^2 \sum_{h=1}^{L} \frac{(1-f_{2h})}{n_{2h}} p_{1h}^2 \hat{S}_{2yh}^2;$$

and

$$a_h \frac{(n_1 - n_{1h})}{n_{2h}(n_1-1)}; \quad f_1 = \frac{n_1}{N}; \quad f_{2h} = \frac{n_{2h}}{n_{1h}};$$

$$\hat{S}_{2yh}^2 = \frac{1}{n_{2h}-1} \sum_{s_{2h}} \left(y_k - \bar{y}_{2h}\right)^2;$$

$$\bar{y}_{2h} = \frac{1}{n_{2h}} \sum_{s_{2h}} y_k$$

and $\bar{y}_{2,st} = \sum_{h=1}^{L} p_{1h} \bar{y}_{2h}$.

Example 2: Let us assume that, for the sampling design described in Example 1, we also have auxiliary data, x_k, available in the first phase s_1. If we assume that the slopes (β_h) vary among the strata, we can assume that the following model $y_k = x_k' \beta_h + \varepsilon_k$ holds, where $E(\varepsilon_k) = 0, E(\varepsilon_k^2) = \sigma_k^2, k \in s_{1h}, h = 1, ..., L$, and $E(\varepsilon_k \varepsilon_\ell) = 0$ for $k \neq \ell$, for $k, \ell \in s_{1h} h = 1, ..., L$. This model gives us a separate regression estimator, that is,

$$\hat{Y}_{SEP,REG} = \sum_{h=1}^{L} \frac{N}{n_1} \frac{n_{1h}}{n_{2h}} \sum_{s_2} g_{2k} y_k$$

where

$$g_{2k} = 1 + \left(\sum_{s_{1h}} x_k' - \sum_{s_{2h}} \frac{n_{1h}}{n_{2h}} x_k'\right) \left(\sum_{s_{2h}} \frac{n_{1h}}{n_{2h}} \frac{x_k x_k'}{\sigma_k^2}\right)^{-1} \frac{x_k}{\sigma_k^2}$$

if $k \in s_{2h}$. In each stratum h, the slopes β_h are estimated as

$$\hat{\boldsymbol{B}}_{2h} = \left(\sum_{s_{2h}} \frac{n_{1h}}{n_{2h}} \frac{x_k x_k'}{\sigma_k^2}\right)^{-1} \left(\sum_{s_{2h}} \frac{n_{1h}}{n_{2h}} \frac{x_k y_k}{\sigma_k^2}\right).$$

The variance of $\hat{Y}_{SEP,REG}$ is estimated as being the sum of the variance components of each phase. These components are $\hat{V}_1$ ($\hat{Y}_{EXP}$) and $\hat{V}_2(\hat{Y}_{SEP,REG})$, where $\hat{V}_1$ ($\hat{Y}_{EXP}$) was defined in example 1. Variance $\hat{V}_2(\hat{Y}_{SEP,REG})$ is obtained by replacing variable y_k by $e_k = g_k(y_k - x_k' \hat{\boldsymbol{B}}_h)$ in $\hat{V}_2(\hat{Y}_{EXP})$. The estimated variance of $\hat{Y}_{SEP,REG}$ is therefore:

$$\hat{V}(\hat{Y}_{SEP,REG}) = \frac{N^2(1-f_1)}{n_1} \sum_{h=1}^{L} p_{1h}\left[(1-a_h)\hat{S}_{2yh}^2 + \frac{n_1}{n_1-1}\left(\bar{y}_{2h} - \bar{y}_{2,st}\right)^2\right] + \sum_{h=1}^{L} \frac{N^2(1-f_{2h})}{n_{2h}} p_{1h}^2 \hat{S}_{2eh}^2$$

where

$$\hat{S}_{2eh}^2 = \sum_{s_{2h}} \frac{(e_k - \bar{e}_h)^2}{n_{2h}-1}$$

and

$$\hat{S}_{2yh}^2 = \frac{1}{n_{2h}-1} \sum_{s_{2h}} \left(y_k - \bar{y}_{2h}\right)^2.$$

5.2 Non-nested Sampling

These two examples are taken from Des Raj (1968, pages 142–149). We are using them to illustrate the results of sections 3 and 4. We consider two different sampling designs.

With the first sampling design, we assume that: (i) the first sample s_1 of size n_1 is selected with a simple random sampling design without replacement from population U; and (ii) the second sample s_2 of size n_2 is selected either by using measurements of size x_i found in the first sample s_1 (nested case) or by selecting it independently (non-nested case) from the first sample s_1 in a manner proportional to

size x_i (known for all units of the population). The resulting estimator is

$$\hat{Y}_{\text{EPTAR}} = \frac{N}{n_1}\frac{\sum_{s_1} x_i}{n_2}\sum_{s_2}\frac{y_i}{x_i}.$$

For the second sampling design, we assume that the two samples s_1 and s_2 have been selected using a simple random sampling design without replacement. Here again, we examine the nested and non-nested cases. We assume that we find the auxiliary observation x_i for any unit selected in the first sample s_1. The estimator is $\hat{Y}_{\text{RAT}} = (N/n_1\Sigma_{s_1}x_i)(\Sigma_{s_2}y_i/\Sigma_{s_2}x_i) = \hat{X}\hat{R}$. Table 3 summarizes these two sampling designs, as well as this corresponding estimators with their estimated variances for the nested and non-nested cases.

The undefined terms in Table 3 are given by $p_{1i} = x_i/\Sigma_{s_1}x_i; p_i = x_i/\Sigma_U x_i; V(\hat{Y}_p) = 1/n_1\Sigma_U p_i(y_i/p_i - Y)^2$; $S_{y-Rx} = (N-1)^{-1}\Sigma_U(y_i - Rx_i)^2$; $f_2 = n_2/N$ $f_1 = n_1/N$, and $R = Y/X$.

Table 3 shows that there is little difference in the variances between the nested and non-nested cases. For $\hat{Y}_{\text{EPTAR}}$, the variance will be smaller for the nested case if the coefficient of variation (CV) of variable y is smaller that that of variable x. For $\hat{Y}_{\text{RAT}}$, the variance will be smaller for the nested case if $\rho\,\text{CV}(\bar{y}) < \text{CV}(\bar{x})$ where ρ is the correlation between y and x.

5.3 Two Statistics Canada Surveys

Several Statistics Canada surveys use double sampling. We will illustrate the ideas presented in this paper using two business surveys. These surveys are the Quarterly Retail Commodity Survey (QRCS) and the Survey of Employment, Payrolls and Hours (SEPH). The Quarterly Retail Commodity Survey uses nested double sampling, whereas the Survey of Employment, Payrolls and Hours (SEPH) uses non-nested double sampling.

The Quarterly Retail Commodity Survey: The purpose of the (QRCS) is to obtain detailed information on retail commodity sales on a quarterly basis. The RCS is a sub-sample of the Monthly Survey of Retail Trade (MRTS), a monthly survey. The MRTS measures mainly sales by trade group (group of three or four-digit codes of the 1980 Standard Industrial Classification (SIC)), by province and for certain census metropolitan areas (CMA). The target population is statistical companies with statistical locations identified on the Business Register and which are active in the retail trade. About 16,000 companies are interviewed each month. The population is stratified by province, territory, certain CMA and by trade group.

The MRTS is stratified in H strata, based on size (2-3 groups), geography (10 provinces, 2 territories) and industry (16 main groups). This sample is restratified independently for the QRCS. The QRCS stratification differs from the MRTS geographically, by size and by industry. A sub-sample is selected using the "new" stratification of the MRTS sample. The QRCS estimate is based on a double-ratio estimator that uses auxiliary data (sales) from the MRTS. The second-phase sampling unit (QRCS) remains the statistical company. The first-phase sample is restratified by trade group, by province and by size based on the most recent information from the MRTS. For stratification purposes, each company is assigned a province and a dominant trade group based on the one that generates the most sales. The two-phase estimator is used by the MRTS. Binder, Babyak, Brodeur, Hidiroglou, and Jocelyn (2000) derived a variance estimator that took into account the sampling design and the estimation method. They expressed variance estimators of the total as simple sums of appropriate residual terms for the case of the ratio estimator.

The results of Binder *et al.* (2000) can be adapted to incorporate the optimal regression estimator in each phase. We assume that the auxiliary information (x_{1k}) is known at

Table 3
Two Sampling Designs with Nested and Non-nested Samples

	Sampling design 1	Sampling design 2
Sampling Design	$N \to n_1$ (SRSWOR)	$N \to n_1$ (SRSWOR)
	$n_1 \to n_2$ (PPSWOR)	$n_1 \to n_2$ (SRSWOR)
Estimator	$\hat{Y}_{\text{EPTAR}} = \frac{N}{n_1}\sum_{s_2}\frac{y_i}{n_2 p_{1i}}$	$\hat{Y}_{\text{RAT}} = \sum_{s_1}\frac{\sum_{s_2} y_i}{\sum_{s_2} x_i} = \hat{X}\hat{R}$
Variance		
Nested	$N^2\frac{(1-f_1)}{n_1}S_y^2 + \frac{V(\hat{Y}_p)}{n_2}$	$\frac{N^2(1-f_1)}{n_1}\left(2RS_{xy} - R^2S_x^2\right) + N^2\frac{(1-f_2)}{n_2}S_{y-Rx}^2$
Non-nested	$N^2\frac{(1-f_1)}{n_1}R^2S_x^2 + \frac{V(\hat{Y}_p)}{n_2}\left[1 + \frac{1}{n_1}(1-f_1)\frac{S_x^2}{\bar{X}^2}\right]$	$\frac{N^2(1-f_1)}{n_1}R^2S_x^2 + N^2\frac{(1-f_2)}{n_2}S_{y-Rx}^2$

the level of population U, either for each unit $k \in U$ or for the total $X_{1k} = \sum_U x_{1k}$. The QRCS sampling design can be formally stated as follows. The population is stratified in H strata U_h; $h = 1, ..., H$, and simple random samples without replacement s_{1h}, of size n_{1h}, are selected in each stratum U_h. The x_k variable is observed for each unit belonging to s_1. The resulting first-phase sample, $s_1 = \cup_{h=1}^{H} s_{1h}$, is then stratified in strata s_{1g}, $g = 1, ..., G$. The stratification of s_1 is independent of the stratification of the universe U. A simple random sample s_{2g} of size n_{2g} is then selected from each stratum s_{1g}, $g = 1, ..., G$. We observe (y_k, x_k'), where $x_k = (x_{1k}', x_{2k}')'$ for each unit belonging to sample $s_2 = \cup_{g=1}^{G} s_{2g}$. We assume that models $y_k = x_{1k}'\beta_1 + \varepsilon_{1k}$ and $y_k = x_k'\beta + \varepsilon_{2k}$ hold for s_1 and s_2 respectively. For each of these models $\varepsilon_{1k} \sim (0, \sigma_1^2 z_{1k})$ and $\varepsilon_{2k} \sim (0, \sigma_2^2 z_{2k})$ where z_{1k} and z_{1k} are known positive factors. If $z_{1k} \neq 1$ or $z_k \neq 1$ for all units $k \in U$, the data can be standardized by dividing them either by $\sqrt{z_{1k}}$ or $\sqrt{z_{2k}}$. The resulting optimal regression estimator for the total Y is given by:

$$\tilde{\bar{Y}}_{OPT} = \hat{\bar{Y}}_{HT} + \left(\bar{X}_1 - \tilde{\bar{X}}_1\right)' \tilde{\bar{B}}_{1,OPT} + \left(\hat{\bar{X}} - \tilde{\bar{X}}\right)' \tilde{\bar{B}}_{OPT}$$

where the components of $\tilde{\bar{Y}}_{OPT}$ were defined in section 3.1. The simplified form (without double sums) of the variance of $\tilde{\bar{Y}}_{OPT}$ is:

$$\hat{V}\left(\tilde{\bar{Y}}_{OPT}\right) = \sum_{h=1}^{H} N_h^2 \left(1 - f_{1h}\right) \frac{\hat{S}_{1h}^2}{n_{1h}} + \sum_{g=1}^{G} n_{1g}^2 \left(1 - f_{2g}\right) \frac{\hat{S}_{2g}^2}{n_{2g}} + \sum_{h=1}^{H} \sum_{g=1}^{G} \frac{N_h^2 \left(1 - f_{1h}\right) n_{2g}^2 \left(1 - f_{2g}\right)}{n_{1h}^2 \left(n_{1h} - 1\right)} \frac{\hat{S}_{2hg}^2}{n_{2h}}$$

where the variances are defined by

$$\hat{S}_{1h}^2 = \frac{1}{n_{1h} - 1} \left\{ \sum_{g=1}^{G} \sum_{k=1}^{n_{2gh}} \frac{n_{1g}}{n_{2g}} \tilde{e}_{1k}^2 - \frac{1}{n_{1h}} \left(\sum_{g=1}^{G} \sum_{k=1}^{n_{2gh}} \frac{n_{1g}}{n_{2g}} \tilde{e}_{1k} \right)^2 \right\};$$

$$\hat{S}_{2hg}^2 = \frac{1}{n_{2hg} - 1} \sum_{k=1}^{n_{2hg}} \left(\tilde{e}_{1k} - \tilde{\bar{e}}_{1(hg)} \right)^2$$

and

$$\hat{S}_{2g} = \frac{1}{n_{2g} - 1} \sum_{k=1}^{n_{2g}} \left(\tilde{e}_{2k} - \tilde{\bar{e}}_{2h} \right)^2.$$

The means in these estimated variances are

$$\tilde{\bar{e}}_{1(hg)} = \frac{1}{n_{2hg}} \sum_{k=1}^{n_{2hg}} \tilde{e}_{1k}, \tilde{\bar{e}}_{1(hg)} = \frac{1}{n_{2hg}} \sum_{k=1}^{n_{2hg}} \tilde{e}_{1k}$$

and

$$\tilde{\bar{e}}_{2h} = \frac{1}{n_{2g}} \sum_{k=1}^{n_{2g}} \tilde{e}_{2k}.$$

Here, n_{2hg} is the number of units selected in sample s_2 belonging to the intersection of strata U_h and s_{1g}. Also, the required residuals are $\tilde{e}_{1k} = g_{1k}(y_k - x_{1k}'\tilde{\bar{B}}_{1,OPT})$ and $\tilde{e}_{2k} = g_{2k}(y_k - x_k'\tilde{\bar{B}}_{OPT})$. The adjustment factors g_{1k} and g_{2k} are as defined in section 4.1.

The Survey of Employment, Payrolls and Hours: The objective of this survey is to obtain estimates of the number of paid employees, the average weekly payroll and other related variables using various combinations of industry and province. This survey was recently redesigned to use administrative data for all businesses included in the survey universe. The survey produces estimates based on both the administrative data (ADMIN sample) and data directly obtained by a survey known as the Business Payroll Survey (BPS).

The ADMIN sample s_1 consists of some 200,000 units selected from universe U_1 of the pay deduction accounts to obtain the administrative data. The sampling design for this sample is stratified Bernoulli (by region), and the sampling rate varies between 10% to 100% amongst the different strata (region). The size of the sample represents approximately 20% of the total number of pay deduction accounts. Only two variables represented as $(x_{1k}^{(1)})$ are available from the administrative source: these are the number of paid employees and the gross monthly payroll.

The BPS sample s_2 consists of approximately 10,000 establishments drawn from the Business Register U_2. The BPS collects the same two variables as the administrative source, namely, the number of paid employees and the gross monthly payroll denoted as $(x_{1k}^{(2)})$, several other variables $(x_{2k}^{(2)})$ of interest defined by type of employee (employees paid by the hour, salaried, active owners, other employees), and variables of interests, such as the number of paid hours and weekly earnings, $(y_k^{(2)})$. More information on the BPS is provided in Rancourt and Hidiroglou (1998).

The BPS is stratified by industry type, geographic region and size (varying from two to three groups based on the number of employees). These strata were designed to take into account the different regression models between $y_k^{(2)}$ and $x_k^{(2)}$. The resulting estimated regression coefficients are used to predict $\hat{y}_k$ for each sampled administrative record. There are two steps involved in the estimation of the total for a given variable of interest. First, the sampling weights $w_k^{(1)}$ associated with the administrative data are calibrated using known regional population counts, N_i, for regions U_{1i}, $i = 1, ..., I$. The adjusted weight of a sample unit k belonging to region U_{1i} is $\tilde{w}_k^{(1)} = w_k^{(1)} g_{1i}$, where $g_{1i} = N_i / \sum_{s_{1i}} w_k^{(1)}$ and $s_{1i} = s_1 \cap U_{1i}$. Second, $y_k^{(2)}$ is regressed on $x_k^{(2)}$ using subsets $s_{2,j}$, $j = 1, ..., J$, of the s_2 sample. The $s_{2,j}$

subsets, classified by industry, region and sometimes size, are formed in advance to obtain the best possible regression fits. For each subset $s_{2,j}$, the estimated regression vectors $\hat{B}_j$ are obtained as:

$$\hat{\hat{B}}_j = \left(\sum_{s_{2,j}} w_k^{(2)} x_k^{(2)} x_k^{\prime(2)} / \hat{\sigma}_k^2\right)^{-1} \sum_{s_{2,j}} w_k^{(2)} x_k^{(2)} y_k^{(2)} / \hat{\sigma}_k^2;$$

$$j = 1, \ldots, J$$

where $w_k^{(2)}$ is the sampling weight for each sampled establishment, and $\hat{\sigma}_k^2$ are known positive factors that control the impact of outliers or define the required estimator. For example, if $\hat{\sigma}_k^2$ is proportional to one of the components of $x_k^{(2)}$, we obtain the ratio estimator. The estimator of total for a variable y is therefore $\hat{Y} = \sum_{j=1}^{J} \sum_{s_{1,j}} \tilde{w}_k^{(1)} x_k^{\prime(1)} \hat{\hat{B}}_j$, where $s_{1,j}$ is a partition of s_1 corresponding to the subsets defining $s_{2,j}$. SEPH is an example of a non-nested double sampling sampling design. More details of the SEPH redesign are available in Hidiroglou (1995) and Hidiroglou, Latouche, Armstrong and Gossen (1995).

6. CONCLUSION

Nested and non-nested double sampling are usually treated separately in the literature. Given that the population total Y is of interest, and that there is auxiliary information available, this paper has unified the estimation procedures for these two sampling methods using an optimal regression approach. Also, for the nested case, the procedure has been linked to the GREG procedure proposed by Hidiroglou and Särndal (1998). For the non-nested case, the method used by Deville (1999) has been extended when there are also auxiliary data at the population level. Lastly, practical examples were provided to illustrate this theory.

REFERENCES

BERGER, Y. (1998). Rate of convergence for asymptotic variance for the Horvitz-Thompson estimator. *Journal of Statistical Planning and Inference*, 74, 149-168.

BINDER, D.A., BABYAK, C., BRODEUR, M., HIDIROGLOU, M.A. and JOCELYN, W. (2000). Variance estimation for two-phase stratified sampling. *The Canadian Journal of Statistics*, 28, 4, 751-764.

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

BREWER, K. (2000). Deriving and estimating an approximate variance for the Horvitz-Thompson estimator using only first order inclusion probabilities. In the *Proceedings of the Second International Conferences on Establishment Surveys*. Buffalo, New York, 1417-1422.

CASSADY, R.J., and VALLIANT, R. (1993). Conditional properties of post-stratified estimation under normal theory. *Survey Methodology*, 19, 183-192.

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

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

DES RAJ (1968). *Sampling Theory*. TMH Edition.

DEVILLE, J.-C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey methodology*, 25, 193-204.

DEVILLE, J.-C. (1999). Simultaneous calibrating of several surveys. *Proceedings: Symposium 1999, Combining Data from Different Sources*, 207-212.

HARTLEY, H.O., and RAO, J.N.K. (1962). Sampling with unequal probabilities and without replacement. *Annals of Mathematical Statistics*, 33, 350-374.

HIDIROGLOU, M.A. (1995). Sampling and estimation for stage one of the canadian survey of employment, payrolls and hours survey redesign. *Proceedings of The Survey Methods Section*, Statistical Society of Canada, 123-128.

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

HIDIROGLOU, M.A., and SÄRNDAL, C.-E. (1998). Use of auxiliary information for two-phase sampling. *Survey Methodology*, 24, 11-20.

KORN, E.L., and GRAUBARD, B.I. (1999). *Analysis of Health Surveys*. Wiley series in probability and Statistics.

MONTANARI, G.E. (1987). Post-sampling efficient prediction in large-scale surveys. *International Statistical Review*, 55, 191-202.

MONTANARI, G.E. (1998). On regression estimation of finite population means. *Survey Methodology*, 24, 69-77.

MONTANARI, G.E. (2000). Conditioning on auxiliary variables means in finite population inference. *Australian New Zealand Journal of Statistics*, 42, 407-421.

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

RANCOURT, E., and HIDIROGLOU, M.A. (1998). Use of administrative records in the Canadian survey of employment, payrolls and hours. *Proceedings of the Survey Methods Section*, 39-47.

RAO, J.N.K. (1973). On double sampling for stratification and analytic surveys . *Biometrika*, 60, 125-133.

RAO, J.N.K. (1994). Estimation of totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics*, 10, 153-166.

RÓSEN, B. (2000). A user's guide to pareto πps sampling. In the *Proceedings of the Second International Conference on Establishment Surveys*, Buffalo, New York, 289-294.

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

SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, Y. (1992). *Model assisted survey sampling*. New York, Springer-Verlag.

TAM, S. M.(1984). On covariances from nested samples. *The American Statistician*. 38, 288-289.

TILLÉ, Y. (2001) *Théorie des Sondages : Échantillonnage et estimation en population finies*. Dumond.

Survey Methodology, December 2001
Vol. 27, No. 2, pp. 155-169
Statistics Canada

Estimation Using the Generalised Weight Share Method: The Case of Record Linkage

PIERRE LAVALLÉE and PIERRE CARON[1]

ABSTRACT

More and more, databases are combined using record linkage methods to increase the amount of available information. When there is no unique identifier to perform the matching, a probabilistic linkage is used. A record on the first file is linked to a record on the second file with a certain probability, and then a decision is made on whether this link is a true link or not. This process usually requires a certain amount of manual resolution that is costly in terms of time and employees. Also, this process often leads to a complex linkage. That is, the linkage between the two databases is not necessarily one-to-one, but can rather be many-to-one, one-to-many, or many-to-many.

Two databases combined using record linkage can be seen as two populations linked together. We consider in this paper the problem of producing estimates for one of the populations (the target population) using a sample selected from the other one. We assume that the two populations have been linked together using probabilistic record linkage. To solve the estimation problem issued from a complex linkage between the population where the sample is selected and the target population, Lavallée (1995) suggested the use of the Generalised Weight Share Method (GWSM). This method is an extension of the Weight Share Method presented by Ernst (1989) in the context of longitudinal household surveys.

The paper will first provide a brief overview of record linkage. Secondly, the GWSM will be described. Thirdly, the GWSM will be adapted to provide three different approaches that take into account linkage weights issued from record linkage. These approaches will be: (1) use all non-zero links with their respective linkage weights; (2) use all non-zero links above a given threshold; and (3) choose the links randomly using Bernoulli trials. For each of the approaches, an unbiased estimator of a total will be presented together with a variance formula. Finally, some simulation results that compare the three proposed approaches to the Classical Approach (where the GWSM is used based on links established through a decision rule) will be presented.

KEY WORDS: Generalised weight share method; Record linkage; Estimation; Clusters.

1. INTRODUCTION

To augment the amount of available information, data from different sources are increasingly being combined. These databases are often combined using record linkage methods. When the files involved have a unique identifier that can be used, the linkage is done directly using the identifier as a matching key. When there is no unique identifier, a probabilistic linkage is used. In that case, a record on the first file is linked to a record on the second file with a certain probability, and then a decision is made on whether this link is a true link or not. Note that this process usually requires a certain amount of manual resolution that is costly in terms of time and employees.

We consider the production of an estimate of a total (or a mean) of one target clustered population when using a sample selected from another population linked to the first population. We assume that the two populations have been linked together using probabilistic record linkage. Note that this type of linkage often leads to a complex linkage between the two populations. That is, the linkage between the units of each of the two populations is not necessarily one-to-one, but can rather be many-to-one, one-to-many, or many-to-many.

To solve the estimation problem caused by a complex linkage between the population where the sample is selected and the target population, Lavallée (1995) suggested the use of the Generalised Weight Share Method (GWSM). This method is an extension of the Weight Share Method presented by Ernst (1989). Although this last method has been developed in the context of longitudinal household surveys, it was shown that the Weight Share Method can be generalised to situations where a target population of clusters is sampled through the use of a frame which refers to a different population, but somehow linked to the first one.

The problem that is considered in this paper is to estimate the total of a characteristic of a target population that is naturally divided into clusters. Assuming that the sample is obtained by the selection of units within clusters, if at least one unit of a cluster is selected, then the whole cluster is interviewed. This usually leads to cost reductions as well as the possibility of producing estimates on the characteristics of both the clusters and the units.

In the present paper, we will try to answer the following questions:

a) Can we use the GWSM to handle the estimation problem related to populations linked together through record linkage?

[1] Pierre Lavallée and Pierre Caron, Statistics Canada, Business Survey Methods Division, Ottawa, Ontario, K1A 0T6, e-mail: plavall@statcan.ca and caropie@statcan.ca.

b) Can we adapt the GWSM to take into account the linkage weights issued from record linkage?

c) Can GWSM help in reducing the manual resolution required by record linkage?

d) If there is more than one approach to use the GWSM, is there a "better" approach?

It will be seen that the answer is clearly yes to (a) and (b). However, for question (c), it will be shown that there is a price to pay in terms of an increase to the sample size, and therefore to the collection costs. For question (d), although there is no definite answer, some approaches seem to generally be more appropriate.

The paper will first provide a brief overview of record linkage. Secondly, the GWSM will be described. Thirdly, the GWSM will be adapted to provide three different approaches that take into account linkage weights issued from record linkage. These approaches will be: (1) use all non-zero links with their respective linkage weights; (2) use all non-zero links above a given threshold; and (3) choose the links randomly using Bernoulli trials. For each of the approaches, an unbiased estimator of a total will be presented together with a variance formula. Finally, some simulation results that compare the three proposed approaches to the Classical Approach (where the GWSM is used based on links established through a decision rule) will be presented.

2. RECORD LINKAGE

The concepts of record linkage were introduced by Newcome, Kennedy, Axford and James (1959) and formalised in the mathematical model of Fellegi and Sunter (1969). As described by Bartlett, Krewski, Wang and Zielinski (1993), record linkage is the process of bringing together two or more separately recorded pieces of information pertaining to the same unit (individual or business). Record linkage is sometimes also called exact matching, in contrast to statistical matching. This last process attempts to link files that have few units in common (see Budd and Radner 1969, Budd 1971, Okner 1972, and Singh, Mantel, Kinack and Rowe 1993). With statistical matching, linkages are based on similar characteristics rather than unique identifying information. In the present paper, we will restrict ourselves to the context of record linkage. However, the developed theory could also be used for statistical matching.

Suppose that we have two files A and B containing characteristics relating to two populations U^A and U^B, respectively. The two populations are somehow related to each other. They can represent, for example, exactly the same population, where each of the files contains a different set of characteristics of the units of that population. They can also represent different populations, but with some natural links between them. For example, one population can be one of parents, and the other population one of children belonging to the parents. Note that the children usually live in households that can be viewed as clusters. Another example is one of an agricultural survey where the first population is a list of farms as determined by the Canadian Census of Agriculture and the second population is a list of taxation records from the Canadian Customs and Revenue Agency (CCRA). In the first population, each farm is identified by a unique identifier called the FarmID and some additional variables such as the name and address of the operators that are collected through the Census questionnaire. The second population consists of taxation records of individuals who have declared some form of agricultural income. These individuals live in households. The unique identifier on those records is either a social insurance number or a corporation number depending on whether or not the business is incorporated. However, each income tax report submitted to CCRA contains similar variables (name and address of respondent, *etc.*) as those collected by the Census.

The purpose of record linkage is to link the records of the two files A and B. If the records contain unique identifiers, then the matching process is trivial. For example, in the agriculture example, if both files would contain the FarmID, the matching process could be done using a simple matching procedure. Unfortunately, often a unique identifier is not available and then the linkage process needs to use some probabilistic approach to decide whether two records of the two files are linked together or not. With this linkage process, the likelihood of a correct match is computed and, based on the magnitude of this likelihood, it is decided whether we have a link or not.

Formally, we consider the product space $\boldsymbol{A}\times\boldsymbol{B}$ from the two files A and B. Let j indicate a record (or unit) from file A (or population U^A) and k a record (or unit) from file B (or population U^B). For each pair (j,k) of $\boldsymbol{A}\times\boldsymbol{B}$, we compute a linkage weight reflecting the degree to which the pair (j,k) is likely to be a true link. The higher the linkage weight is, the more likely the pair (j,k) is a true link. The linkage weight is commonly based on the ratios of the conditional probabilities of having a match μ and an unmatch $\bar{\mu}$ given the result of the outcome of the comparison C_{qjk} of the characteristic q of the records j from A and k from B, $q = 1, ..., Q$. That is,

$$\dot{\theta}_{jk} = \log_2\left\{\frac{P(\mu_{jk}|C_{1jk}C_{2jk}\ldots C_{Qjk})}{P(\bar{\mu}_{jk}|C_{1jk}C_{2jk}\ldots C_{Qjk})}\right\}$$

$$= \dot{\theta}_{1jk} + \dot{\theta}_{2jk} + \ldots + \dot{\theta}_{Qjk} + \dot{\theta}_{\bullet jk} \qquad (2.1)$$

where $\dot{\theta}_{qjk} = \log_2\left\{\frac{P(C_{qjk}|\mu_{jk})}{P(C_{qjk}|\bar{\mu}_{jk})}\right\}$ for $q = 1, ..., Q$, and

$\dot{\theta}_{\bullet jk} = \log_2\left\{\frac{P(\mu_{jk})}{P(\bar{\mu}_{jk})}\right\}$.

The mathematical model proposed by Fellegi and Sunter (1969) takes into account the probabilities of an error in the linkage of units j from A and k from B. The linkage weight is then defined as

$$\theta_{jk}^{FS} = \sum_{q=1}^{Q} \theta_{qjk}^{FS}$$

where

$$\theta_{jk}^{FS} = \begin{cases} \log_2 & \text{if characteristic } q \text{ of pair } (jk) \text{ agrees} \\ \log_2 ((1-\eta_{qjk})/(1-\bar{\eta}_{qjk})) & \text{otherwise} \end{cases}$$

with $\eta_{qjk}^{FS} = P(\text{characteristic } q \text{ agrees } | \mu_{jk})$ and $\bar{\eta}_{qjk} = P(\text{characteristic } q \text{ agrees } | \bar{\mu}_{jk})$. Note that the definition of θ_{jk}^{FS} assumes that the Q comparisons are independent.

The linkage weights given by (2.1) are defined on $\boldsymbol{R}$, the set of real numbers, *i.e.*, $\theta_{jk} \in]-\infty, +\infty[$. When the ratio of the conditional probabilities of having a match μ and an unmatch $\bar{\mu}$ is equal to 1, we get $\theta_{jk} = 0$. When this ratio is close to 0, θ_{jk} tends to $-\infty$. It might then be more convenient to define the linkage weights on $[0, +\infty[$. This can be achieved by taking the antilogarithm of θ_{jk}. We then obtain the following linkage weight θ_{jk}:

$$\theta_{jk} = \frac{P(\mu_{jk} | C_{1jk} C_{2jk} \ldots C_{Qjk})}{P(\bar{\mu}_{jk} | C_{1jk} C_{2jk} \ldots C_{Qjk})}. \qquad (2.2)$$

Note that the linkage weight θ_{jk} is equal to 0 when the conditional probabilities of having a match μ is equal to 0. In other words, we have $\theta_{jk} = 0$ when the probability of having a true link for (j, ik) is nul.

Once a linkage weight θ_{jk} has been computed for each pair (j, k) of $\boldsymbol{A} \times \boldsymbol{B}$, we need to decide whether the linkage weight is sufficiently large to consider the pair (j, k) a link. This is typically done using a decision rule. With the approach of Fellegi and Sunter, we use an upper threshold θ_{High} and a lower threshold θ_{Low} to which each linkage weight θ_{jk} is compared. The decision is made as follows:

$$D(j,k) = \begin{cases} \text{link} & \text{if } \theta_{jk} \geq \theta_{\text{High}} \\ \text{can be a link} & \text{if } \theta_{\text{Low}} < \theta_{jk} < \theta_{\text{High}} \\ \text{nonlink} & \text{if } \theta_{jk} \leq \theta_{\text{Low}}. \end{cases} \qquad (2.3)$$

The lower and upper thresholds θ_{Low} and θ_{High} are determined by *a priori* error bounds based on false links and false nonlinks. When applying decision rule (2.3), some clerical decisions are needed for those linkage weights falling between the lower and upper thresholds. This is generally done by looking at the data, and also by using auxiliary information. In the agriculture example, variables such as date of birth, street address and postal code, which are available on both sources of data, can be used for this purpose. By being automated and also by working on a probabilistic basis, some errors can be introduced in the record linkage process. This has been discussed in several papers, namely Bartlett *et al.* (1993), Belin (1993) and Winkler (1995).

The application of decision rule (2.3) leads to the definition of an indicator variable $l_{jk} = 1$ if the pair (j, k) is considered to be a link, and 0 otherwise. As for the decisions that need to be taken for those linkage weights falling between the lower and upper thresholds, some manual intervention may be needed to decide on the validity of the links. In the case where the files A and B represent the same population (with a different set of characteristics), it is likely that for each unit j from file A, there will be only one unit linked in file B. That is, the units should be linked on a one-to-one basis. Note that decision rule (2.3) does not prevent the existence of many-to-one, one-to-many, or many-to-many links. As mentioned before, because of the probabilistic aspect of the record linkage process, which might introduce some errors, there could be more than one link per unit. In practice, this problem is usually solved by some manual intervention. In the agriculture example, it can occur that multiple operators of a farm each submit a tax report to CCRA for the same farm (one-to-many). Similarly, an operator who runs more than one farm could submit only one income tax report for his operations (many-to-one). Finally, one can imagine a scenario of many-to-many links when an operator runs more than one farm, where each farm has a number of different operators. These situations can be represented by Figure 1. In Figure 1, unit j=1 of U^A has a one-to-one link to unit k=1 of U^B; unit j=2 forms to a one-to-many link to units k=2 and k=4; and units j=2 and j=3 together form a many-to-one link to unit k=4. For the agriculture example, it is clear that deciding on the validity of the links is more difficult than the case of the same population since the former allows the possibility of having true many-to-one or one-to-many situations.

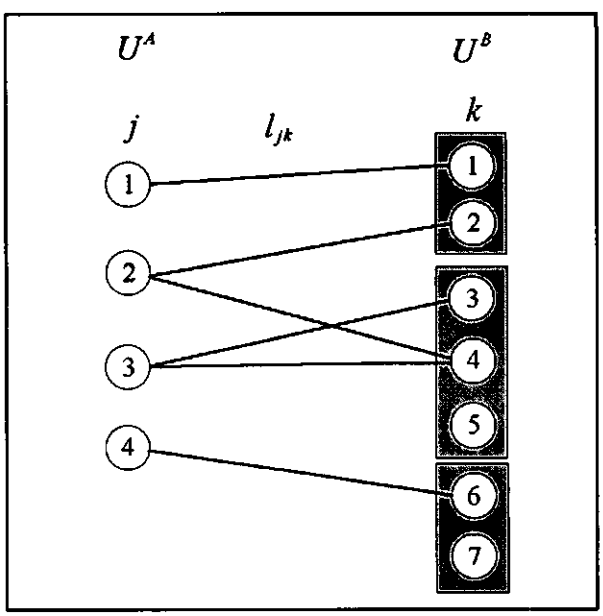

Figure 1. Example of links

3. THE GENERALISED WEIGHT SHARE METHOD

The GWSM is described in Lavallée (1995). It is an extension of the Weight Share Method described by Ernst (1989) but in the context of longitudinal household surveys. Various implications of using the Weight Share Method for longitudinal household surveys have been described by Gailly and Lavallée (1993). The GWSM can be viewed as a generalisation of *Network Sampling* and also of *Adaptive Cluster Sampling*. These two sampling methods are described in Thompson (1992), and Thompson and Seber (1996).

Suppose that a sample s^A of m^A units is selected from the population U^A of M^A units using some sampling design. Let π_j^A be the selection probability of unit j. We assume $\pi_j^A > 0$ for all $j \in U^A$.

Let the population U^B contain M^B units. This population is divided into N clusters where cluster i contains M_i^B units. For example, in the context of social surveys, the clusters can be households and the units can be the persons within the households. For business surveys, the clusters can be enterprises and the units can be the establishments within the enterprises. For the agriculture example, the clusters can be households, and the units, persons within the household who file an income tax report to CCRA.

We suppose that there exists a link between the units j of population U^A and the units k of clusters i of the population U^B. This link is identified by an indicator variable $l_{j,ik}$, where $l_{j,ik} = 1$ if there exists a link between unit $j \in U^A$ and unit $ik \in U^B$, and 0 otherwise. Note that there might be some units j of population U^A for which there is no link with any unit k of a cluster i of population U^B, *i.e.*, $L_j^A = \sum_{i=1}^{N} \sum_{k=1}^{M_i^B} l_{j,ik} = 0$ for all $j \in U^A$. Also, there can be zero, one or more links for any unit k of a cluster i of population U^B, *i.e.*, $L_{ik} = \sum_{j=1}^{M^A} l_{j,ik} = 0, L_{ik} = 1$ or $L_{ik} > 1$ for any $k \in U^B$.

With the GWSM, we have the following constraint:

> Each cluster i of U^B must have at least one link with a unit j of U^A, *i.e.*, $L_i = \sum_{j=1}^{M^A} \sum_{k=1}^{M_i^B} l_{j,ik} > 0$.

This constraint is essential for the GWSM to produce unbiased estimates. We will see in section 4 that in the context of record linkage, this constraint might not be satisfied.

For each unit j selected in s^A, we identify the units ik of U^B that have a non-zero link with j, *i.e.*, $l_{j,ik} = 1$. For each identified unit ik, we suppose that we can establish the list of the M_i^B units of cluster i containing this unit. Then, each cluster i represents by itself a population U_i^B where $U^B = \bigcup_{i=1}^{N} U_i^B$. Let Ω^B be the set of the n clusters identified by the units $j \in s^A$.

From population U^B, we are interested in estimating the total $Y^B = \sum_{i=1}^{N} \sum_{k=1}^{M_i^B} y_{ik}$ for some characteristic y. An important constraint that is imposed in the measurement (or interviewing) process of y is to consider all units within the same cluster. That is, if a unit is selected in the sample, then every unit of the cluster containing the selected unit is interviewed. This constraint is one that often arises in surveys for two reasons: cost reductions and the need for producing estimates on clusters. As an example, for social surveys, there is normally a small marginal cost for interviewing all persons within the household. On the other hand, household estimates are often of interest with respect to poverty measures, for example. For the agriculture example, one value of interest is the total farm revenue per household. In that case, we need to interview all persons within the household.

By using the GWSM, we want to assign an estimation weight w_{ik} to each unit k of an interviewed cluster i. To estimate the total Y^B belonging to population U^B, one can then use the estimator

$$\hat{Y} = \sum_{i=1}^{n} \sum_{k=1}^{M_i^B} w_{ik} y_{ik} \tag{3.1}$$

where n is the number of interviewed clusters and w_{ik} is the weight attached to unit k of cluster i. With the GWSM, the estimation process uses the sample s^A together with the links existing between U^A and U^B to estimate the total Y^B. The links are in fact used as a bridge to go from population U^A to population U^B, and vice versa.

The GWSM allocates to each interviewed unit ik a final weight established from an average of weights calculated within each cluster i entering into $\hat{Y}$. An *initial weight* that corresponds to the inverse of the selection probability is first obtained for all units k of cluster i of $\hat{Y}$ having a non-zero link with a unit $j \in s^A$. An initial weight of zero is assigned to units not having a link. The *final weight* is obtained by calculating the ratio of the sum of the initial weights for the cluster over the total number of links for that cluster. This final weight is finally assigned to all units within the cluster. Note that the fact of allocating the same estimation weight to all units has the considerable advantage of ensuring consistency of estimates for units and clusters.

Formally, each unit k of cluster i entering into $\hat{Y}$ is assigned an initial weight w'_{ik} as follows:

$$w'_{ik} = \sum_{j=1}^{M^A} l_{j,ik} \frac{t_j}{\pi_j^A} \tag{3.2}$$

where $t_j = 1$ if $j \in s^A$ and 0 otherwise. Note that a unit ik having no link with any unit j of U^A has automatically an initial weight of zero. The final weight w_i is given by

$$w_i = \frac{\sum_{k=1}^{M_i^B} w'_{ik}}{\sum_{k=1}^{M_i^B} L_{ik}} \tag{3.3}$$

where $L_{ik} = \sum_{j=1}^{M^A} l_{j,ik}$. The quantity L_{ik} represents the number of links between the units of U^A and the unit k of cluster i of U^B. The quantity $L_i = \sum_{k=1}^{M_i^B} L_{ik}$ then corresponds to the total number of links present in cluster i. Finally, we assign $w_{ik} = w_i$ for all $k \in U_i^B$ and use equation (3.1) to estimate the total Y^B.

Using this last expression, it was shown in Lavallée (1995) that the GWSM is design unbiased. Further, let $z_{ik} = Y_i/L_i$ for all $k \in i$, where $Y_i = \sum_{k=1}^{M_i^B} y_{ik}$. Then, $\hat{Y}$ can be expressed as

$$\hat{Y} = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} \sum_{i=1}^{N} \sum_{k=1}^{M_i^B} l_{j,ik} z_{ik} = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} Z_j \qquad (3.4)$$

and the variance of $\hat{Y}$ is given by

$$\mathrm{Var}(\hat{Y}) = \sum_{j=1}^{M^A} \sum_{j'=1}^{M^A} \frac{(\pi_{jj'}^A - \pi_j^A \pi_{j'}^A)}{\pi_j^A \pi_{j'}^A} Z_j Z_{j'} \qquad (3.5)$$

where $\pi_{jj'}^A$ is the joint probability of selecting units j and j'. See Särndal, Swensson and Wretman (1992) for the calculation of $\pi_{jj'}^A$ under various sampling designs. The variance $\mathrm{Var}(\hat{Y})$ may be unbiasedly estimated from the following equation:

$$\mathrm{Var}(\hat{Y}) = \sum_{j=1}^{M^A} \sum_{j'=1}^{M^A} \frac{(\pi_{jj'}^A - \pi_j^A \pi_{j'}^A)}{\pi_{jj'}^A \pi_j^A \pi_{j'}^A} t_j Z_j t_{j'} Z_{j'}. \qquad (3.6)$$

Another unbiased estimator of the variance $\mathrm{Var}(\hat{Y})$ may be developed in the form of Yates and Grundy (1953).

In presenting the Weight Share Method in the context of longitudinal surveys, Ernst (1989) proposed the use of constants α in the definition of the estimation weights. In the general context of the GWSM, the use of the same type of constants can be proposed. Let us define $\alpha_{j,ik} \geq 0$ for all pairs (j, ik), with $\alpha_i = \sum_{j=1}^{M^A} \sum_{k=1}^{M_i^B} \alpha_{j,ik} = 1$. We can then obtain new estimation weights as follows. For each unit k of cluster i entering into $\hat{Y}$, assign the following initial weight $w_{ik}^{\prime\alpha}$:

$$w_{ik}^{\prime\alpha} = \sum_{j=1}^{M^A} \alpha_{j,ik} \frac{t_j}{\pi_j^A}. \qquad (3.7)$$

The final weight w_i^α is given by

$$w_i^\alpha = \sum_{k=1}^{M_i^B} w_{ik}^{\prime\alpha} = \sum_{k=1}^{M_i^B} \sum_{j=1}^{M^A} \alpha_{j,ik} \frac{t_j}{\pi_j^A}. \qquad (3.8)$$

Finally, we assign $w_{ik}^\alpha = w_i^\alpha$ for all $k \in U_i^B$ and use equation (3.1) to estimate the total Y^B.

In the context of longitudinal surveys, Ernst (1989) noted that the most common choice for the constants α is the one where each individual receives one of two values: 0, or a non-zero value that is equal for all the remaining units within the cluster. In the present context, this would mean to let $\alpha_{j,ik} = 0$ for all j and k in a subset U_i^{0B} of U_i^B, say, and $\alpha_{j,ik}$ = constant for all j and k in the complement subset $U_i^{\bar{0}B}$. Back to the context of longitudinal surveys, Kalton and Brick (1995) looked at the determination of optimal values for the α of Ernst (1989) where the optimality is measured in terms of minimal variance. They concluded that: "in the two-household case, the equal household weighting scheme minimises the variance of the household weights around the inverse selection probability weight when the initial sample is an equal epsem (equal probability) one." They also added that "in the case of an approximately epsem sample, the equal household weighting scheme should be close to the optimal, at least for the case where the members of the household at time t come from one or two households at the initial wave." This suggests that, for the GWSM, the choice of letting the constants α being 0 for some units and a positive value that is equal for all the remaining units within the cluster should be close to the optimal.

4. THE GWSM AND RECORD LINKAGE

With record linkage, the links $l_{j,ik}$ are established between files A and B, or population U^A and population U^B, using a probabilistic process. As mentioned before, record linkage uses a decision rule D such as (2.3) to decide whether there is a link or not between unit j from file A and unit ik from file B. Once the links are established, we then have the two populations U^A and U^B linked together, with the links identified by the indicator variable $l_{j,ik}$. Note that the decision rule (2.3) does not prevent the existence of complex links (many-to-one, one-to-many, or many-to-many).

Although the links can be complex, the GWSM can be used to estimate the total Y^B from population U^B using a sample s^A obtained from population U^A. Therefore, the answer is yes to question (a) stated in the introduction. Note that the estimates produced by the application of the GWSM might however not be unbiased if the constraint mentioned in section 3 is not satisfied. In that case, the use of the estimation weight (3.3) underestimates the total Y^B. To solve this problem, one practical solution is to collapse two clusters in order to get at least one non-zero link $l_{j,ik}$ for cluster i. This solution usually requires some manual intervention. Another solution is to impute a link by choosing one link at random within the cluster, or to choose the link with the largest linkage weight $\theta_{j,ik}$. Note that it might also happen that for a unit j of U^A, there is no non-zero link $l_{j,ik}$ with any unit ik of U^B. This is however not a problem since the only coverage in which we are interested is the one of U^B.

It is now clear that the GWSM can be used in the context of record linkage. The GWSM with the populations U^A and U^B linked together using record linkage with the decision rule (2.3) will be referred to as the Classical Approach.

Now, with the Classical Approach, the use of the GWSM is based on links identified by the indicator variable $l_{j,ik}$. Is it necessary to establish whether there is positively a link for each pair (j, ik), or not? Would it be easier to simply use the linkage weights $\theta_{j,ik}$ (without using any decision rule) to estimate the total Y^B from U^B using a sample from U^A? These questions lead to question (b) on whether or not it is possible to adapt the GWSM to take into account the linkage weights θ issued from record linkage.

In the present section, we will see that the answer to question (b) is yes by providing three approaches where the GWSM uses the linkage weights θ. The first approach is to use all the non-zero links identified through the record linkage process, together with their respective linkage weights θ. The second approach is the one where we use all the non-zero links with linkage weights above a given threshold θ_{High}. The third approach is one where the links are randomly chosen with probabilities proportional to the linkage weights θ.

4.1 Approach 1: Using all Non-Zero Links With Their Respective Linkage Weights

When using all non-zero links with the GWSM, one might want to give more importance to links that have large linkage weights θ, compared to those that have small linkage weights. By definition, for each pair (j, ik) of $A \times B$, the linkage weight $\theta_{j,ik}$ reflects the degree to which the pair (j, ik) is likely to be a true link. We then no longer use the indicator variable $l_{j,ik}$ identifying whether there is a link or not between unit j from U^A and unit k of cluster i from U^B. Instead, we use the linkage weight $\theta_{j,ik}$ obtained in the first steps of the record linkage process. (This assumes that the file with the linkage weights is available. In practice, the only available file is often the linked file obtained at the end of the linkage process, once some manual resolution has been performed. In this case, the linkage weights are no longer available and the three proposed approaches to be used with the GWSM are immaterial to reduce the problem of manual resolution). Note that by doing so, we do not need any decision to be taken to establish whether there is a link or not between two units.

For each unit j selected in s^A, we identify the units ik of U^B that have a non-zero linkage weight with unit j, *i.e.*, $\theta_{j,ik}>0$. Let $\Omega^{RL,B}$ be the set of the n^{RL} clusters identified by the units $j \in s^A$, where "RL" stands for "Record Linkage". Note that because we use all non-zero linkage weights, we have $n^{RL} \geq n$. We now obtain the initial weight w_{ik}^{*RL} by directly replacing the indicator variable l in equations (3.2) and (3.3) by the linkage weight θ.

$$w_{ik}^{*RL} = \sum_{j=1}^{M^A} \theta_{j,ik} \frac{t_j}{\pi_j^A}. \tag{4.1}$$

The final weight w_i^{RL} is given by

$$w_i^{RL} = \frac{\sum_{k=1}^{M_i^B} w_{ik}^{*RL}}{\sum_{k=1}^{M_i^B} \Theta_{ik}} \tag{4.2}$$

where $\Theta_{ik} = \sum_{j=1}^{M^A} \theta_{j,ik}$. Finally, we assign $w_{ik}^{RL} = w_i^{RL}$ for all $k \in U_i^B$. Note that by being present both at the numerator and denominator of equation (4.2), the linkage weights $\theta_{j,ik}$ do not need to be between 0 and 1. They just need to represent the relative likelihood of having a link between two units from populations U^A and U^B. It is also interesting to note that by letting $\alpha_{j,ik} = \theta_{j,ik} / \Theta_i$ where $\Theta_i = \sum_{j=1}^{M^A} \sum_{k=1}^{M_i^B} \theta_{j,ik}$, we obtain, for the estimation weight w_i^{RL}, an equivalent formulation to the one given by (3.7) and (3.8).

With the Classical Approach, we stated the constraint that each cluster i of U^B must have at least one link with a unit j of U^A, *i.e.*, $L_i = \sum_{j=1}^{M^A} \sum_{k=1}^{M_i^B} l_{j,ik} > 0$. This constraint is translated here into the need of having for each cluster i of U^B at least one non-zero linkage weight $\theta_{j,ik}$ with a unit j of U^A, *i.e.*, $\theta_i = \sum_{j=1}^{M^A} \sum_{k=1}^{M_i^B} \theta_{j,ik} > 0$. In theory, the record linkage process does not insure that this constraint is satisfied. It might then turn out that for a cluster i of U^B, there is no non-zero linkage weight $\theta_{j,ik}$ with any unit j of U^A. In that case, the use of the estimation weight (4.2) underestimates the total Y^B. To solve this problem, the same solutions proposed in the context of the indicator variables $l_{j,ik}$ can be used. That is, a solution is to collapse two clusters in order to get at least one non-zero linkage weight $\theta_{j,ik}$. Unfortunately, this solution might require some manual intervention, which has been avoided up to now by not using the decision rule (2.3). A better solution is to impute a link by choosing one link at random within the cluster, and then assign arbitrarily a small value for $\theta_{j,ik}$ to the chosen link (for example, the smallest calculated non-zero linkage weight).

To estimate the total Y^B belonging to population U^B, one can use the estimator

$$\hat{Y}^{RL} = \sum_{i=1}^{n^{RL}} \sum_{k=1}^{M_i^B} w_{ik}^{RL} y_{ik}. \tag{4.3}$$

Following the same steps used to obtain equation (3.4), one can write $\hat{Y}^{RL}$ as

$$\begin{aligned} \hat{Y}^{RL} &= \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} \sum_{i=1}^{N} \sum_{k=1}^{M_i^B} \theta_{j,ik} z_{ik}^{RL} \\ &= \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} Z_j^{RL} \end{aligned} \tag{4.4}$$

where $z_{ik}^{RL} = Y_i / \Theta_i$ for all $k \in U_i^B$, and $\Theta_i = \sum_{k=1}^{M_i^B} \Theta_{ik} = \sum_{k=1}^{M_i^B} \theta_{j,ik}$. Using this last expression, it can be shown that $\hat{Y}^{RL}$ is design unbiased for Y^B. The variance of $\hat{Y}^{RL}$ is given by

$$\text{Var}(\hat{Y}^{\text{RL}}) = \sum_{j=1}^{M^A}\sum_{j'=1}^{M^A} \frac{(\pi^A_{jj'} - \pi^A_j \pi^A_{j'})}{\pi^A_j \pi^A_{j'}} Z_j^{\text{RL}} Z_{j'}^{\text{RL}}. \quad (4.5)$$

4.2 Approach 2: Use all Non-Zero Links Above a given Threshold

Using all non-zero links with the GWSM as in Approach 1 might require the manipulation of large files of size $M^A \times M^B$. This is because it might turn out that most of the records between files A and B have non-zero linkage weights θ. In practice, even if this happens, we can expect that most of these linkage weights will be relatively small or negligible to the extent that, although non-zero, the links are very unlikely to be true links. In that case, it might be useful to only consider the links with a linkage weight θ above a given threshold θ_{High}.

For this second approach, we again no longer use the indicator variable $l_{j,ik}$ identifying whether there is a link or not, but instead, we use the linkage weight $\theta_{j,ik}$ that are above the threshold θ_{High}. The linkage weights below the threshold are considered as zeros. We therefore define the linkage weight:

$$\theta^T_{j,ik} = \begin{cases} \theta_{j,ik} \text{ if } \theta_{j,ik} \geq \theta_{\text{High}} \\ 0 \quad \text{otherwise}. \end{cases}$$

For each unit j selected in s^A, we identify the units ik of U^B that have $\theta^T_{j,ik} > 0$. Let $\Omega^{\text{RLT},B}$ be the set of the n^{RLT} clusters identified by the units $j \in s^A$, where "RLT" stands for "Record Linkage with Threshold". Note that $n^{\text{RLT}} \leq n^{\text{RL}}$. On the other hand, we have $n^{\text{RLT}} = n$ if the record linkage between U^A and U^B is done by using the decision rule (2.3) with $\theta_{\text{High}} = \theta_{\text{Low}}$.

The initial weight $w^{*\text{RLT}}_{ik}$ is given by

$$w^{*\text{RL}}_{ik} = \sum_{j=1}^{M^A} \theta^T_{j,ik} \frac{t_j}{\pi^A_j}. \quad (4.6)$$

The final weight w_i^{RLT} is given by

$$w_i^{\text{RL}} = \frac{\sum_{k=1}^{M_i^B} w^{*\text{RLT}}_{ik}}{\sum_{k=1}^{M_i^B} \Theta^T_{ik}} \quad (4.7)$$

where $\Theta^T_{ik} = \sum_{j=1}^{M^A} \theta^T_{j,ik}$. Finally, we assign $w^{\text{RLT}}_{ik} = w_i^{\text{RLT}}$ for all $k \in U_i^B$. As for Approach 1, it is interesting to note that by letting $\alpha_{j,ik} = \theta^T_{j,ik} / \Theta^{T,B}_i$ where $\Theta^{T,B}_i = \sum_{j=1}^{M^A} \sum_{k=1}^{M_i^B} \theta^T_{j,ik}$, we obtain, for the estimation weight w_i^{RLT}, an equivalent formulation to the one given by (3.7) and (3.8).

The number of zero linkage weights θ^T will be greater than or equal to the number of zero linkage weights θ used by Approach 1. Therefore, the constraint that each cluster i of U^B must have at least one non-zero linkage weight $\theta^T_{j,ik}$ with a unit j of U^A might be more difficult to satisfy. In that case, the use of the estimation weight (4.7) underestimate the total Y^B. To solve this problem, the same solutions proposed before can be used.

To estimate the total Y^B, one can use the same estimator as (4.3), where we replace the number of identified clusters n^{RL} by n^{RLT}, and the estimation weight w_{ik}^{RL} by w_{ik}^{RL}. As for estimator (4.3), it can be shown that this estimator $\hat{Y}^{\text{RLT}}$ is design unbiased.

4.3 Approach 3: Choose the Links by Random Selection

In order to avoid making a decision on whether there is a link or not between unit j from U^A and unit k of cluster i from U^B, one can decide to simply choose the links at random from the set of non-zero links. For this, it is reasonable to choose the links with probabilities proportional to the linkage weights θ. This can be achieved by Bernoulli trials where, for each pair (j, ik), we decide on accepting a link or not by generating a random number $u_{j,ik} \sim U(0,1)$ that is compared to a quantity proportional to the linkage weight $\theta_{j,ik}$.

In the point of view of record linkage, this approach cannot be considered as optimal. When using the decision rule (2.3) of Fellegi and Sunter, the idea is to try to minimise the number false links and false nonlinks. The link $l_{j,ik}$ is accepted only if the linkage weight $\theta_{j,ik}$ is large enough (*i.e.*, $\theta_{j,ik} \geq \theta_{\text{High}}$), or if it is moderately large (*i.e.*, $\theta_{\text{Low}} < \theta_{jk} < \theta_{\text{High}}$) and has been accepted after manual resolution. Selecting the links randomly using Bernoulli trials might lead to the selection of links that would have not been accepted through the decision rule (2.3), even though the selection probabilities are proportional to the linkage weights. Some of the resulting links between the two populations U^A and U^B might then be false ones, and some units that are not linked might be false nonlinks. The linkage errors are therefore likely to be higher than if the decision rule (2.3) would be used. However, in the present context, the quality of the linkage is of secondary interest. The present problem is to try to estimate the total Y^B using the sample s^A selected from U^A, and not to evaluate the quality of the links. The precision of the estimates of Y^B will in fact be measured only in terms of the sampling variability of the estimators, by conditioning on the linkage weights $\theta_{j,ik}$. Note that this sampling variability will take into account the random selection of the links, but not the linkage errors.

The first step before performing the Bernoulli trials is to transform the linkage weights in order to restrict them to the [0,1] interval. By looking at (2.1), it can be seen that the linkage weights $\theta_{j,ik}$ correspond in fact to a logit transformation (in base 2) of the probability $P(\mu_{jk} \mid C_{1jk} C_{2jk} \ldots C_{Qjk})$. Similarly, the linkage weights given by (2.2) depend only on this probability. Hence, one way to transform the linkage weights is simply to use the

probability $P(\mu_{jk} \mid C_{1jk} C_{2jk} \ldots C_{Qjk})$. From (2.1), we obtain this result by using the function $\tilde{\theta} = 2^{\theta}/(1+2^{\theta})$. From (2.2), we use $\tilde{\theta} = \theta/(1+\theta)$. When the linkage weights are not obtained through (2.1) nor (2.2), a possible transformation is to divide each linkage weight by the maximum possible value $\theta_{Max} = \max_{j=1,\, i=1,\, k=1}^{M^A, N, M_i^B} \theta_{j,ik}$. Note that we assume that the linkages weights are all greater than or equal to zero, which is the case with definition (2.2), but not necessarily in general.

Once the adjusted linkage weights $\tilde{\theta}_{j,ik}$ have been obtained, for each pair (j, ik), we generate a random number $u_{j,ik} \sim U(0,1)$. Then, we set the indicator variable θ_{Hig} to 1 if $u_{j,ik} \leq \tilde{\theta}_{j,ik}$, and 0 otherwise. This process provides a set of links similar to the ones used in the Classical Approach, with the exception that now the links have been determined randomly instead of through a decision process comparable to (2.3). Note that since $E(\tilde{l}_{j,ik}) = \tilde{\theta}_{j,ik}$, the sum of the adjusted linkage weights $\tilde{\theta}_{j,ik}$ corresponds to the expected total number of links L from the Bernoulli process in, $\boldsymbol{A} \times \boldsymbol{B}$, *i.e.*,

$$\sum_{j=1}^{M^A} \sum_{i=1}^{N} \sum_{k=1}^{M_i^B} \tilde{\theta}_{j,ik} = L. \tag{4.8}$$

For each unit j selected in s^A, we identify the units ik of U^B that have $\tilde{l}_{j,ik} = 1$. Let $\tilde{\Omega}^B$ be the set of the $\tilde{n}$ clusters identified by the units $j \in s^A$. Note that $\tilde{n} \leq n^{RL}$. Unfortunately, in contrast to n^{RL} and n^{RLT}, the random number of clusters $\tilde{n}$ is hardly comparable to n.

The initial weight $\tilde{w}'_{ik}$ is defined as follows:

$$\hat{Y} = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} \sum_{i=1}^{N} \sum_{k=1}^{M_i^B} l_{j,ik} z_{ik} = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} Z_j. \tag{4.9}$$

The final weight $\tilde{w}_{ik}$ is given by

$$\tilde{w}_{ik} \frac{\sum_{k=1}^{M_i^B} \tilde{w}'_{ik}}{\sum_{k=1}^{M_i^B} \tilde{L}_{ik}} \tag{4.10}$$

where $\tilde{L}_{ik} = \sum_{j=1}^{M^A} \tilde{l}_{j,ik}$. The quantity $\tilde{L}_{ik}$ represents the realised number of links between the units of U^A and the unit k of cluster i of population U^B. Finally, we assign $\tilde{w}_{ik} = \tilde{w}_i$ for all $k \in U_i^B$.

To estimate the total Y^B, we can use the estimator

$$\hat{\tilde{Y}} = \sum_{i=1}^{\tilde{n}} \sum_{k=1}^{M_i^B} \tilde{w}_{ik} y_{ik}. \tag{4.11}$$

By conditioning on the accepted links $\tilde{l}$, it can be shown that estimator (4.11) is conditionally design unbiased and hence, unconditionally design unbiased. Note that by conditioning on $\tilde{l}$, the estimator (4.11) is then equivalent to (3.1). To get the variance of $\hat{\tilde{Y}}$, again conditional arguments need to be used. Letting the subscript 1 indicate that the expectation is taken over all possible sets of links, we have

$$\mathrm{Var}(\hat{\tilde{Y}}) = E_1 \mathrm{Var}_2(\hat{\tilde{Y}}) + \mathrm{Var}_1 E_2(\hat{\tilde{Y}}). \tag{4.12}$$

First, from conditional unbiasedness, we have

$$E_2(\hat{\tilde{Y}}) = Y^B. \tag{4.13}$$

Therefore,

$$\mathrm{Var}_1 E_2(\hat{\tilde{Y}}) = 0. \tag{4.14}$$

Second, from (3.5), we directly have

$$\mathrm{Var}_2(\hat{\tilde{Y}}) = \sum_{j=1}^{M^A} \sum_{j'=1}^{M^A} \frac{(\pi_{jj'}^A - \pi_j^A \pi_{j'}^A)}{\pi_j^A \pi_{j'}^A} \tilde{Z}_j \tilde{Z}_{j'} \tag{4.15}$$

where $\tilde{Z}_j$ is defined as in (3.4) but with the links l replaced by $\tilde{l}$. Hence, the variance of $\hat{\tilde{Y}}$ can be expressed as

$$\mathrm{Var}_2(\hat{\tilde{Y}}) = E_1\left(\sum_{j=1}^{M^A} \sum_{j'=1}^{M^A} \frac{(\pi_{jj'}^A - \pi_j^A \pi_{j'}^A)}{\pi_j^A \pi_{j'}^A} \tilde{Z}_j \tilde{Z}_{j'} \right) \tag{4.16}$$

where the expectation is taken over all possible sets of links.

With the GWSM, we stated in section 3 a constraint that must be satisfied for unbiasedness of the GWSM. In the present approach, by randomly selecting the links, it is very likely that this constraint will not be satisfied. To solve this problem, we can impute a link by choosing the one with the highest non-zero linkage weight $\theta_{j,ik}$ within the cluster. If there is still no link because all $\theta_{j,ik} = 0$, it is possible to choose one link at random within the cluster. It should be noted that this solution preserves the design unbiasedness of the GWSM.

4.4 Some Remarks

The three proposed approaches do not use the decision rule (2.3). They also not make use of any manual resolution. Hence, the answer to the question (c) of the introduction is yes. That is, GWSM can help in reducing the manual resolution required by record linkage. Note that there is however a price to pay for avoiding manual resolution.

First, with Approach 1, the number n^{RL} of clusters identified by the units $j \in s^A$ is greater than or equal to the number n of clusters identified by the Classical Approach, *i.e.*, when the decision rule (2.3) is used to identify the links. This is because we use all non-zero links, and not just the ones satisfying the decision rule (2.3). As a consequence, the collection costs with Approach 1 will be greater than or equal to the ones related to the use of the Classical Approach. It needs then to be checked which ones are the most important: the collections costs or the costs of manual resolution. Note that if the precision resulting from the use of Approach 1 is much higher than one from the Classical

Approach, it might be more of interest to use the former than the latter.

With Approach 2, we have $n^{RLT} \leq n^{RL}$ and therefore the collection costs of this approach are less than or equal to the ones of Approach 1. If the precision of Approach 2 is comparable to the one of Approach 1, then the former will certainly be more advantageous than the latter. By comparing Approach 2 with the Classical Approach, it can be seen that the collection costs can be almost equivalent if the value of the threshold θ_{High} is chosen to be close to the lower and upper thresholds of the decision rule (2.3). Note that Approach 2 is not using any manual resolution. If the precision of Approach 2 is at least comparable to the one of the Classical Approach, then Approach 2 will have a clear advantage. Note also that if $\theta_{High} = \theta_{Low}$, the two approach differs only in the definition of the estimation weights obtained by the GWSM. Approach 2 uses the linkage weights θ, while the Classical Approach uses the indicator variables l. After setting $\theta_{High} = \theta_{Low}$, it is certainly of interest to verify which approach has the highest precision.

With Approach 3, the number of selected links will be less than or equal to the number of non-zero links used by Approach 1, *i.e.*, $\tilde{n} \leq n^{RL}$. Hence, the collection costs of Approach 3 will be less than or equal to the ones of Approach 1. In terms of precision, it is not clear which variance is likely to be the smallest between to two approaches. As mentioned before, in opposite to n^{RL} and n^{RLT} the random number of clusters $\tilde{n}$ is hardly comparable to n. The two depends on different parameters: The Classical Approach depends on the thresholds θ_{Low} and θ_{High}, while Approach 3 depends on the adjusted linkage weights $\tilde{\theta}_{j,ik}$ that correspond to the selection probabilities of the links.

5. SIMULATION STUDY

A simulation study was performed to evaluate the proposed approaches against the Classical Approach where the decision rule (2.3) is used to determine the links. This study was made by comparing the precision obtained for the estimation of a total Y^B using five different approaches:

Approach 1: use all non-zero links with their respective linkage weights

Approach 2: use all non-zero links above a threshold

Approach 3: choose the links randomly using Bernoulli trials

Approach 4: Classical Approach

Approach 5: use all non-zero links, but with the indicator variable l

Approach 5 is a mixture of Approach 1 and the Classical Approach. It is basically to first accept as links all pairs (j, ik) with a non-zero linkage weights, *i.e.*, assign $l_{j,ik} = 1$ for all pairs (j, ik) where $\theta_{j,ik} > 0$, and 0 otherwise. The GWSM described in section 3 is then used to produce the estimate of Y^B. Approach 5 was added to the simulations to see the effect of using the indicator variable l instead of the linkage weight θ when using all non-zero links. As for the other approaches, Approach 5 can be shown to be unbiased.

Given that all five approaches yield design unbiased estimates of the total Y^B, the quantity of interest for comparing the various approaches was the standard error of the estimate, or simply the coefficient of variation (*i.e.*, the ratio of the square root of the variance to the expected value).

The simulation study was performed based on the agriculture example mentioned throughout the paper. This example corresponds in fact to a real situation occurring at Statistics Canada related to the construction of the Whole Farm Data Base (see Statistics Canada 2000). Note that although the simulation study was based on a real situation, some of the numbers used have been changed for confidentiality reasons. Also, the linkage process did not reflect the exact procedure used within Statistics Canada. For more information on the exact procedure, see Lim (2000). It was felt that these changes do not negate the results of the simulation study. The main purpose of the simulations was to evaluate the proposed approaches against the Classical Approach. It was not intended to solve the problems related to the construction of the Whole Farm Data Base, which could be considered as a secondary goal.

Recall that the agriculture example is one of an agricultural survey where the first population U^A is a list of farms as determined by the Canadian Census of Agriculture. This list is from the 1996 Farm Register, which is essentially a list of all records collected during the 1991 Census of Agriculture with all the updates that have occurred since 1991. It contains a farm operator identifier together with some socio-demographic variables related to the farm operators. The second population U^B is a list of taxation records from the CCRA. This second list is the 1996 Unincorporated CCRA Tax File that contains data on tax filers declaring at least one farming income. It contains a household identifier (only on a sample basis), a tax filer identifier, and also socio-demographic variables related to the tax filers.

At Statistics Canada, Agriculture Division produces estimates on crops and livestocks using samples selected from the Farm Register (population U^A). To create the Whole Farm Data Base, it is of interest to collect tax data for the farms that have been selected in the samples from the Farm Register. This is done by first merging the Farm Register with the Unincorporated CCRA Tax File (population U^B) and then obtaining the tax data from CCRA. As mentioned before, it turns out that the relationship between the farm operators of the Farm Register and the tax filers from the Unincorporated CCRA Tax File is not one-to-one. This is why the GWSM turns out to be a useful approach for producing estimation weights for the tax filers selected through the sample of farm operators from the Farm Register.

Some might argue that there is no need to obtain a set of clusters identified by the units $j \in s^A$, since the target population U^B is one of tax filers from the Unincorporated CCRA Tax File, which is usually available on a census basis. Note however that this is not totally true. Not all variables of interest are available on this file and Statistics Canada needs to pay for the extra variables requested from CCRA. Also, the data from the Unincorporated CCRA Tax File are not free of errors due to keying, coding, *etc.*, and therefore there are some costs related to cleaning up the data. For these reasons, it is found preferable to restrict the data from the target population U^B to a subset only. Since this needs to be done, one way of identifying the set of clusters to be used in the estimate of Y^B is simply to do it through the sample s^A selected from U^A.

Apart from the Classical Approach, all approaches consider the linkage itself between U^A and U^B as a secondary goal, the first one being to produce an estimate Y^B for the target population U^B. However, the application mentioned here is one related to the Whole Farm Data Base, which aims to be an integrated data base. Not having a linkage of good quality between the populations U^A and U^B would lead to erroneous microdata analyses between the crops and livestocks variables measured in the sample s^A and the tax data obtained from U^B. On this aspect, the authors agree that the proposed approaches, with the exception of the Classical Approach, are not viable in the present context. This is true however in a long term point of view. Because manual resolution is needed when using a decision rule such as (2.3), one could suggest to use the proposed approaches to produce some of the required estimates from U^B in the short term, before the final linkage is available, after manual resolution. Recall that the main purpose of the simulations is to evaluate the proposed approaches against the Classical Approach. The agriculture example has not been chosen because it corresponds to a real situation, but more because of the availability of the data. It could have been any other example such as the other one mentioned in the introduction where U^A is a population of parents and U^B a population of children belonging to the parents.

For the purpose of the simulations, two provinces of Canada were considered: New Brunswick and Québec. The former can be considered as a small province and the latter a large one. Table 1 provides the size of the different files. Because the household identifier is not available for the entire population U^B, for the purpose of the simulations, it has been constructed based on a sample. This sample has the household identifier coded for each tax filer. For the non-sample tax filers, the household identifiers were randomly assigned such that the household sizes correspond to the same proportions of household sizes found in the sample.

Table 1
Agriculture Example

	Québec	New Brunswick
Size of Farm Register (U^A)	43017	4930
Size of Tax File (U^B)	52394	5155
Total number of households of U^B	22387	2194
Total number of Non-zero Linkage Weights	105113	13787

The linkage process used for the simulations was a match using five variables. It was performed using the MERGE statement in SAS®. All records on both files were compared to one another in order to see if a potential match had occurred. The record linkage was performed using the following five key variables common to both sources:

- first name (modified using NYSIIS)
- last name (modified using NYSIIS)
- birth date
- street address
- postal code

The first name and last name variables were modified using the NYSIIS system. This basically changes the name in phonetic expressions, which in turn increases the chance of finding matches by reducing the probability that a good match is rejected because of a spelling mistake. For more details about NYSIIS, see Lynch and Arends (1977).

Records that matched on all 5 variables received the highest linkage weight (θ=60). Records that matched on only a subset of at least 2 of the 5 variables received a lower linkage weight (as low as θ=2). It should be noted that the levels of the linkage weights were chosen arbitrarily. As mentioned before, it is not really the levels themselves that are important, but rather the relative importance of the linkage weights between each other.

Records that did not match on any combination of key variables were not considered as potential links, which is equivalent as having a linkage weight of zero. Two different thresholds were used for the simulations: $\theta_{High} = \theta_{Low} = 15$ and $\theta_{High} = \theta_{Low} = 30$. The upper and lower thresholds, θ_{High} and θ_{Low}, were set to be the same to avoid the grey area where some manual intervention is needed when applying the decision rule (2.3).

Note that the constraint related to the use of the GWSM needed to be satisfied. When for a cluster i of U^B there was no non-zero linkage weight $\theta_{j,ik}$ between any units k of this cluster and the units from U^A, we imputed a link by choosing the link with the largest linkage weight $\theta_{j,ik}$ within the cluster. Note that it also happened that for some units j of U^A, there was no non-zero linkage weight $\theta_{j,ik}$ with any unit ik of U^B, this was not considered a problem since the only coverage in which we are interested is the one of U^B. Table 1 provides the total number of non-zero links found in each of the two provinces.

For the simulations, we have selected the sample from U^A (*i.e.*, the Farm Register) using Simple Random Sampling Without Replacement (SRSWOR), without any stratification. We also considered two sampling fractions: 30% and 70%. The quantity of interest Y^B to be estimated was the Total Farming Income.

Since we have the whole population of farms and taxation records, it was possible for us to calculate the theoretical variance for these estimates. It was also possible to estimate this variance by selecting a large number of samples (*i.e.*, performing a Monte-Carlo study), estimating the parameter Y^B for each sample, and then calculating the variance of all the estimates. Both approaches were used. For the simulations, 500 simple random samples were selected for each approach for the two different sampling fractions (30% and 70%). The two thresholds (15 and 30) were also used to better understand the properties of the given estimators.

Because we assumed SRSWOR, the theoretical formulas given in section 4 could be simplified. For example, under SRSWOR, the variance formula (4.5) reduced to the following:

$$\text{Var}(\hat{Y}^{\text{RL}}) = M^A \frac{(1-f)}{f} S^2_{Z,\text{RL}} \qquad (5.1)$$

where $f = m^A/M^A$ is the sampling fraction, $S^2_{Z,\text{RL}} = 1/M^A - 1 \sum_{j=1}^{M^A} (Z_j^{\text{RL}} - \bar{Z}^{\text{RL}})^2$ and $\bar{Z}^{\text{RL}} = 1/M^A \sum_{j=1}^{M^A} Z_j^{\text{RL}}$.

The Monte-Carlo study involved 500 replicates. For each of the two sampling fractions (30% and 70%), 500 simple random samples t were selected, and the expectation and variance for each of the five approaches were then estimated using

$$\hat{E}(\hat{Y}) = \frac{1}{500} \sum_{t=1}^{500} \hat{Y}_t \qquad (5.2)$$

and

$$\hat{V}(\hat{Y}) = \frac{1}{500} \sum_{t=1}^{500} (\hat{Y}_t - \hat{E}(\hat{Y}))^2. \qquad (5.3)$$

The estimated coefficients of variation (CVs) were obtained by using

$$C\hat{V}(\hat{Y}) = 100 \times \frac{\sqrt{\hat{V}(\hat{Y})}}{\hat{E}(\hat{Y})}. \qquad (5.4)$$

The Monte-Carlo process was performed to verify empirically the exactness of the theoretical formulas provided in section 4. The results indicate that all the theoretical formulas provided were exact.

The results of the study are presented in Figures 2.1 to 2.4, Table 2, and Figure 3. Figures 2.1 to 2.4 provide bar charts of the CVs obtained for each of the five approaches. The bar charts are given for the eight cases obtained by crossing the two provinces Québec and New Brunswick, the two sampling fractions 30% and 70%, and the two thresholds 15 and 30. On each bar of the charts, one can find the number of non-zero links between U^A and U^B for each of the five approaches. Note that for Approach 3, it corresponds in fact to the expected number of non-zero links. The number of (expected) non-zero links does not change from one sampling fraction to another. Table 2 shows the average number of clusters interviewed by approach, for each of the eight cases, where the average is taken over the 500 samples used for the simulations. The numbers in parenthesis are the standard deviations. They are relatively small compared to the averages and therefore the number of clusters identified through the sample s^A does not fluctuate greatly from one sample s^A to another. Figure 3 provides scattered plots of the obtained CVs by the average number of clusters identified through the sample s^A, for each of the eight cases.

By looking at the Figures 2.1 to 2.4, it can be seen that in all cases, Approach 1 and Approach 5 provided the smallest CVs for the estimation of the Total Farming Income. Therefore, using all non-zero links yield the greatest precision. Note however that by looking at Table 2, we can see that these approaches also lead to the highest number of clusters identified through the sample selected from U^A. In fact, we can see that the greater the number of clusters used in the estimation is, the greater the precision of the resulting estimates is. This result is shown in Figure 3 where we can see that the CVs tend to decrease as the average number of clusters identified through s^A increases. Although this result is well known in the classical sampling theory, it was not guaranteed to hold in the context of the GWSM. As we can see from equation (3.5), it is not the sample size of s^A that increases, but rather the homogeneity of the derived variables Z_j.

Now, by comparing Approach 1 and Approach 5, it can be seen that the latter always provided the smallest variance. Therefore, this suggests to use the indicator variable l instead of the linkage weight θ when using all non-zero links. Note that it seems this can be generalised since the same phenomenon occurred with Approach 2 and Approach 4 (Classical Approach). Recall that, because $\theta_{\text{High}} = \theta_{\text{Low}}$, the two approaches differ only in the definition of the estimation weights obtained by the GWSM. Approach 2 uses the linkage weights θ, while the Classical Approach uses the indicator variables l. Note that this results goes along the conclusions of Kalton and Brick (1995) since the optimal choice of letting the constants α being 0 for some units and a positive value that is equal for all the remaining units within the cluster corresponds to the use of the indicator variable l.

We now concentrate on Approach 3. For seven out of the eight histograms of Figures 2.1 to 2.4, Approach 3 produced the highest CVs. The only lower CV was obtained for Québec, with the sampling fraction of 30% and the threshold $\theta_{\text{High}} = 30$. It should however be noted that this approach is the one that used the lowest number of non-zero links, and also the lowest average number of clusters identified through s^A. Therefore, this result is not totally surprising. Recall that the number of non-zero links

used by Approach 3 does not depend on the threshold θ_{High} and thus the CVs obtained for Québec with f=0.3 were equal for θ_{High}=15 and θ_{High}=30. For θ_{High}=15, the CV obtained for Approach 3 for Québec was higher than the ones for Approaches 2 and 4, and these two were using more non-zero links, and more clusters. For θ_{High}=30 the CV obtained for Approach 3 was lower than the ones from approaches 2 and 4, but these two were still using more non-zero links, and more clusters. Therefore, there are intermediate situations where with $15 < \theta_{High} < 30$, we should get equal CVs for approaches 3 and 2, and approaches 3 and 4. As a consequence, to get equal CVs between Approach 3 and each of approaches 2 and 4, more non-zero links and more clusters must be used by the latter. This suggests that in some cases, Approach 3 might be more appropriate to use than approaches 2 and 4 because estimates with the same precision can be obtained with lower collection costs.

In order to better compare Approach 3 to the approaches 2 and 4, we forced the number of expected non-zero links to be the same as the number of non-zero links used by approaches 2 and 4. For this, we have transformed the linkage weights $\theta_{j,ik}$ to $\tilde{\tilde{\theta}}_{j,ik}$ in order to have

$$\sum_{j=1}^{M^A} \sum_{i=1}^{N} \sum_{k=1}^{M_i^B} \tilde{\tilde{\theta}}_{j,ik} = L_0 \qquad (5.5)$$

where L_0 is the desired number of non-zero links. The transformation used was

$$\tilde{\tilde{\theta}}_{j,ik} = \begin{cases} \theta_{j,ik}/\theta_{\bullet} & \text{if } \dfrac{\theta_{j,ik}}{\theta_{\bullet}} \le 1 \\ 1 & \text{otherwise} \end{cases} \qquad (5.6)$$

where $\theta_{\bullet}$ was determined iteratively such that (5.5) is satisfied. The use of Approach 3 with the transformation (5.6) is referred to as Approach 6. The results of the simulations are presented in Figures 4.1 to 4.4. As we can see, Approach 6 turned out to have the smallest CVs for half of the cases. For the other cases, Approach 4 yielded the best precision. Note that this situation did not occur for a particular province only, nor a particular sampling fraction, and also nor for a particular threshold. It would therefore be difficult in practice to determine in advance which of Approach 6 or Approach 4 would produce the smallest CVs. Because of this, and because of the fact that Approach 6 (and Approach 3) can produce large linkage errors, Approach 4 should be preferred.

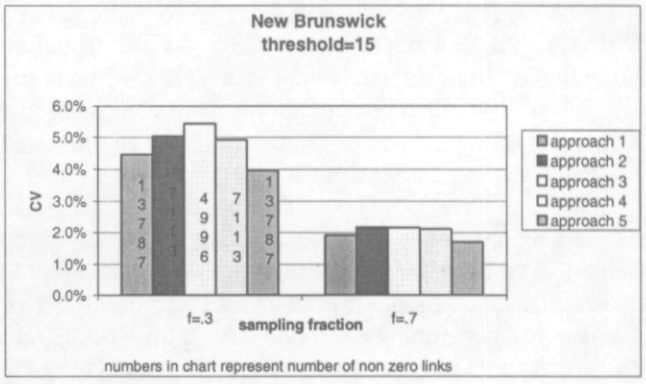

Figure 2.1 CVs for New Brunswick (with $\theta_{high} = \theta_{Low} = 15.$)

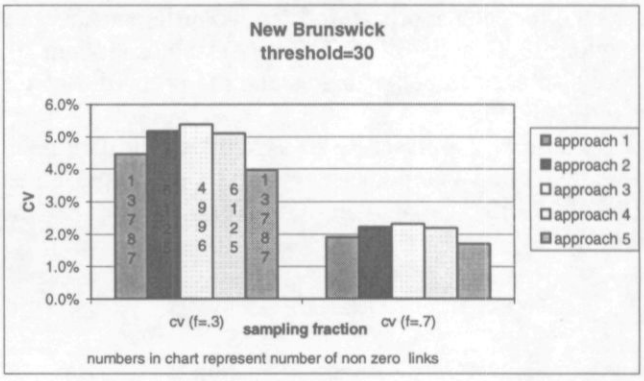

Figure 2.2 CVs for New Brunswick (with $\theta_{High} = \theta_{Low} = 30$)

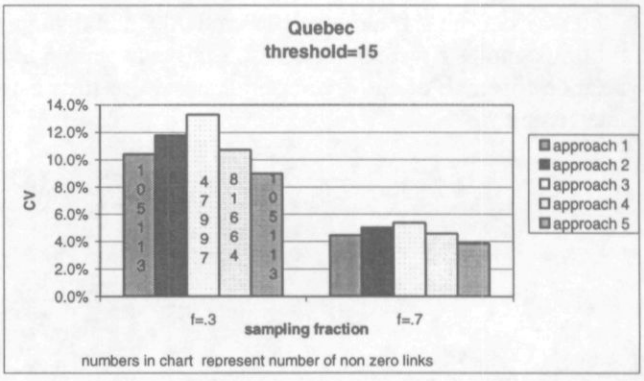

Figure 2.3 CVs for Québec (with $\theta_{High} = \theta_{Low} = 15$)

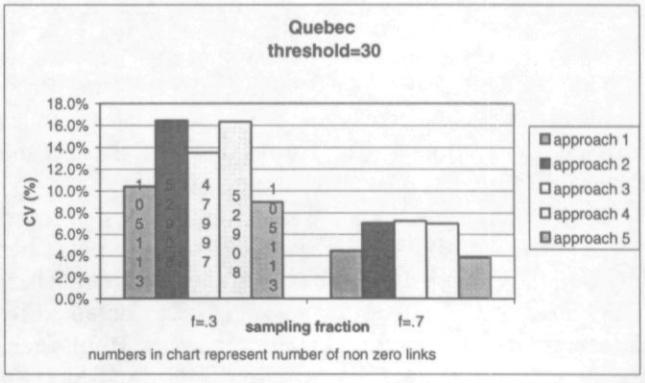

Figure 2.4 CVs for Québec (with $\theta_{High} = \theta_{Low} = 30$)

Table 2
Average Number of Identified Cluster

Threshold	Approach	Average number of identified clusters (s.e.)			
		Quebec		New Brunswick	
		f=.3	f=.7	f=.3	f=.7
15	1	15752(58)	21106(30)	1709(18)	2100(7)
	2	14281(49)	20593(34)	1310(17)	1966(13)
	3	10930(50)	18881(47)	1123(14)	1869(14)
	4	14281(49)	20593(34)	1310(17)	1966(13)
	5	15752(58)	21106(30)	1709(18)	2100(7)
30	1	15752(58)	21106(30)	1709(18)	2100(7)
	2	11310(45)	19139(37)	1215(17)	1924(15)
	3	10930(50)	18881(47)	1123(14)	1869(14)
	4	11310(45)	19139(37)	1215(17)	1924(15)
	5	15752(58)	21106(30)	1709(18)	2100(7)

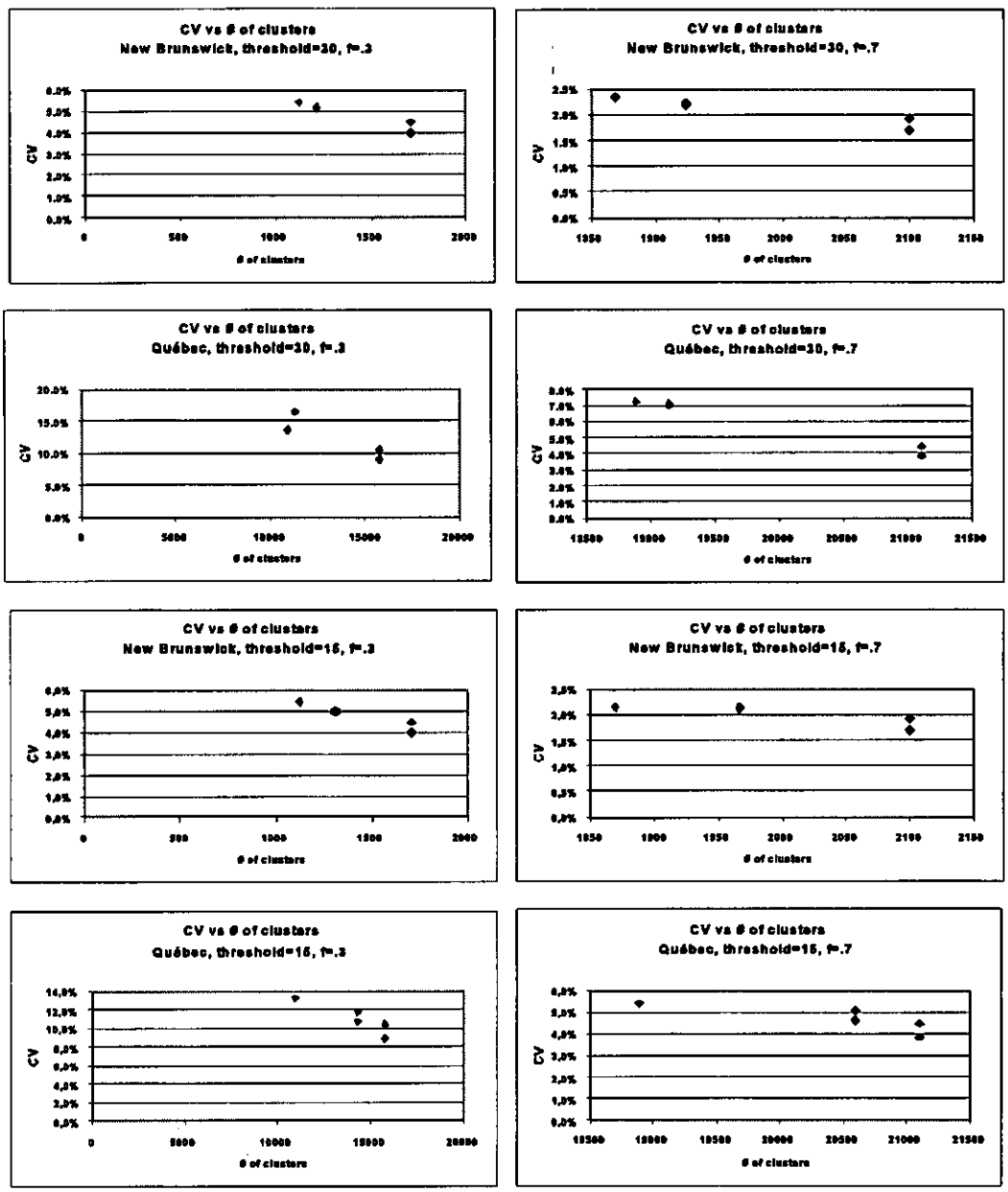

Figure 3. Graphs of CVs versus Average Number of Identified Clusters

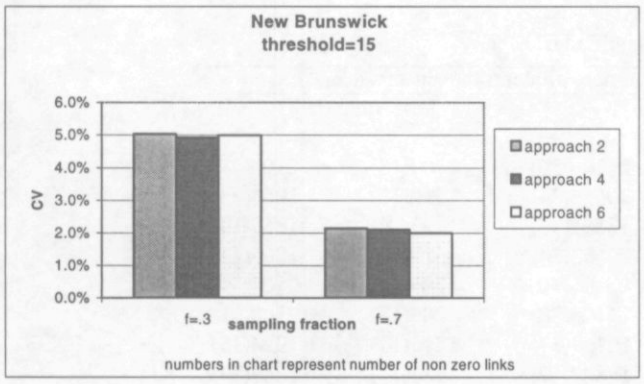

Figure 4.1. CVs for New Brunswick (with $\theta_{High} = \theta_{Low} = 15$).

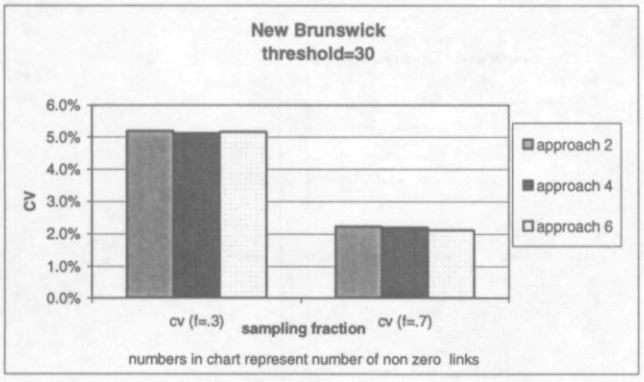

Figure 4.2. CVs for Québec (with $\theta_{High} = \theta_{Low} = 30$).

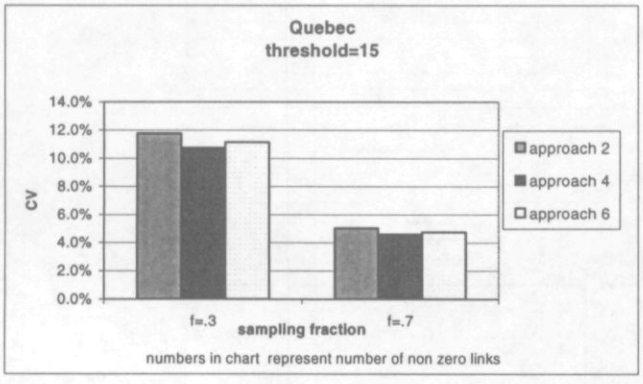

Figure 4.3. CVs for Québec (with $\theta_{High} = \theta_{Low} = 15$).

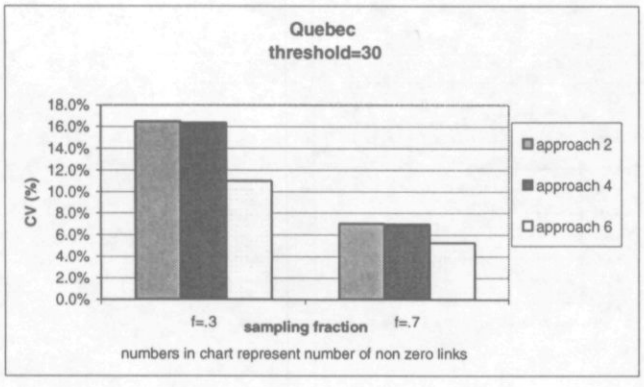

Figure 4.4. CVs for Québec (with $\theta_{High} = \theta_{Low} = 30$).

6. CONCLUSION

In the present paper, we have seen that the GWSM is adaptable to populations linked through Record Linkage. This is in fact simply a natural extension of the case where the links are either present or absent, which corresponds to the use of an indicator variable $l_{j,ik} = 1$ if the pair (j, ik) is considered to be a link, 0 otherwise. When two populations are linked through record linkage, there is always some uncertainty left because the decisions on the links are made using a probabilistic approach. Therefore, replacing the indicator variable $l_{j,ik}$ by the linkage weight $\theta_{j,ik}$ that has been computed for each pair (j, ik) simply makes the GWSM more generalised.

Some simulations were performed using the 1996 Farm Register (population U^A) and the 1996 Unincorporated CCRA Tax File (population U^B). We compared the variances obtained for each of the five approaches: (1) use all non-zero links; (2) use all non-zero links above a threshold; (3) choose links randomly using Bernoulli trials (4) Classical Approach; (5) use all non-zero links, but with the indicator variable l. All results showed that Approach 1 and Approach 5 provide the smallest CVs for the estimation of the Total Farming Income. These two approaches use however the highest number of links, and also the highest number of clusters identified through s^A, which implies the highest collection costs. Because of this, the approaches 2, 3 and 4 might be viewed as good compromises.

For a given threshold θ_{High}, it is preferable to use the indicator variable l instead of the linkage weights θ in the construction of the estimation weights with the GWSM. This result holds even for $\theta_{High} = 0$ (*i.e.*, no threshold is used), as for approaches 1 and 5. The estimates produced with the indicator variable l always had the smallest CVs and this result goes along the conclusions of Kalton and Brick (1995). Hence, Approach 5 should be preferred to Approach 1, and Approach 4 should be preferred to Approach 2.

The use of the threshold θ_{High} is useful to reduce the number of non-zero links to be manipulated. By reducing the number of non-zero links, we reduce as well the number of clusters identified through the sample s^A, and hence we reduce the collection costs associated to the measurement of the variable of interest y within the clusters. Note that by reducing the number of links, we decrease the precision of the estimates produced. Therefore, a choice needs to be made between the desired precision and the collection costs.

The reduction of the number of non-zero links can also be achieved by using the decision rule (2.3) with the two thresholds θ_{Low} and θ_{High}. This decreases the collection costs, but introduces the need of some manual resolution when the linkage weights θ are between θ_{Low} and θ_{High}. The manual resolution leads however to better links, *i.e.*, with less linkage errors. If manual resolution is used only to make the links one-to-one between population U^A and

population U^B, then it might not be necessary since the GWSM is particularly appropriate to handle estimation in situations where the links between U^A and U^B are complex.

When compared to approaches 2 and 4, Approach 3 turned out to be preferable in some cases. Because it would be difficult in practice to determine in advance which of Approach 3 or Approach 4 would produce the smallest CVs, and because of the fact that Approach 3 can produce large linkage errors, Approach 4 should be preferred. Hence, the Classical Approach of using the GWSM with the indicator variable l with links determined using a decision rule such as (2.3) seems the most appropriate approach to estimate the total Y^B using a sample selected from U^A.

ACKNOWLEDGEMENTS

The authors would like to thank the Associate Editor and the two referees for their useful suggestions and comments. These have contributed to improve significantly the quality of the paper.

REFERENCES

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

BELIN, T.R. (1993). Evaluation of sources of variation in record linkage through a factorial experiment. *Survey Methodology*, 19, 13-29.

BUDD, E.C. (1971). The creation of a microdata file for estimating the size distribution of income. *The Review of Income and Wealth*, 17, 317-333.

BUDD, E.C., and RADNER, D.B. (1969). The OBE size distributions series: methods and tentative results for 1964. *American Economic Review*, Papers and Proceedings, LIX, 435-449.

ERNST, L. (1989). Weighting issues for longitudinal household and family estimates. In *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: John Wiley and Sons, 135-159.

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

GAILLY, B., and LAVALLÉE, P. (1993). Insérer des nouveaux membres dans un panel longitudinal de ménages et d'individus: simulations. CEPS/Instead, Document PSELL No. 54, Luxembourg.

KALTON, G., and BRICK, J.M. (1995). Weighting schemes for household panel surveys. *Survey Methodology*, 21, 33-44.

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

LIM, A. (2000). Results of the Linkage between the 1998 Taxation Data and the 1998 Farm Register. Internal document of the Business Survey Methods Division, Statistics Canada.

LYNCH, B.T., and ARENDS, W.L. (1977). Selection of a Surname Coding Procedure for the SRS Record Linkage System. Document of the Sample Survey Research Branch, Statistical Reporting Service, U.S. Department of Agriculture, Washington, D.C.

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

OKNER, B.A. (1972). Constructing a new data base from existing microdata sets: The 1966 merge file. *Annals of Economic and Social Measurement*, 1, 325-342.

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

SINGH, A.C., MANTEL, A.J., KINACK, M.D. and ROWE, G. (1993). Statistical matching: Use of auxiliary information as an alternative to the conditional independence assumption. *Survey Methodology*, 19, 59-79.

STATISTICS CANADA (2000). *Whole Farm Database reference manual*. Publication No. 21F0005GIE, Statistics Canada, 100 pages.

THOMPSON, S.K. (1992). *Sampling*. New York: John Wiley and Sons.

THOMPSON, S.K., and SEBER, G.A. (1996). *Adaptive Sampling*. New York: John Wiley and Sons.

WINKLER, W.E. (1995). Matching and record linkage. *Business Survey Methods*, (Eds. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge and P.S. Kott), New York: John Wiley and Sons, 355-384.

YATES, F., and GRUNDY, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society B*, 15, 235-261.

Survey Methodology, December 2001
Vol. 27, No. 2, pp. 171-181
Statistics Canada

Cross-sectional Estimation in Multiple-Panel Household Surveys

TAKIS MERKOURIS[1]

ABSTRACT

This paper presents weighting procedures that combine information from multiple panels of a repeated panel household survey for cross-sectional estimation. The non static nature of a repeated panel survey is discussed in relation to estimation of population parameters at any wave of the survey. A repeated panel survey with overlapping panels is described as a special type of multiple frame survey, with the frames of the panels forming a time sequence. The paper proposes weighting strategies suitable for various multiple-panel survey situations. The proposed weighting schemes involve an adjustment of weights in domains of the combined panel sample that represent identical time periods covered by the individual panels. A weight adjustment procedure that deals with changes in the panels over time is discussed. The integration of the various weight adjustments required for cross-sectional estimation in a repeated panel household survey is also discussed.

KEY WORDS: Repeated panel surveys; Multiple frames; Temporal domains; Combined panels; Cross-sectional weighting; Weight share method.

1. INTRODUCTION

A panel survey collects the survey data for the same sample elements at different time points (the survey waves). A repeated panel survey is made up of a series of panel surveys, each having fixed duration, with the panels selected at different time points. In a repeated panel household survey a sample of households is selected for each panel from the population of households existing at the start of the panel. Depending on the objectives of the panel survey, one or all individuals in the sampled households become panel members to be followed throughout the duration of the panel or until they leave the survey population. At a subsequent survey wave the household sample consists of all the households in which panel members reside. A review of various types of panel surveys is given in Kalton and Citro (1993). A formalization of related concepts can be found in Deville (1998).

The type of repeated panel household survey considered in this paper consists of two or more panels covering overlapping time periods. A typical example of such a survey is the Canadian Survey of Labour and Income Dynamics (SLID), which employs two overlapping panels of duration of six years each; for a description of the SLID see Lavigne and Michaud (1998). In the SLID, each new panel is introduced three years after the introduction of the previous one. The sample for each panel is made up of two rotation groups from the Canadian Labour Force Survey, which uses a stratified multistage design with an area frame wherein dwellings containing households are the final sampling units.

A panel survey, though primarily conducted for longitudinal purposes, may also be used to produce cross-sectional estimates of population parameters for any survey wave. For cross-sectional purposes, data are usually collected at each survey wave for all individuals living in households that contain at least one selected member. The process of obtaining cross-sectional estimates at any wave of a panel household survey after the first wave presents difficulties arising from the population and panel dynamics. Weighting schemes that deal with dynamic features of a single panel, such as movers and "cohabitants," have been discussed in the literature; see Kalton and Brick (1995), and Lavallée (1995) for details. Yet, there seems to be a paucity of work in the literature on cross-sectional estimation for repeated panel household surveys with overlapping panels; some initial work in the context of the SLID can be found in Lavallée (1994). The cross-sectional estimation problem in such multiple panel surveys is a proper combination of the panels that would account for the changes in the population and in the panels over time.

This paper describes procedures for cross-sectional estimation that combine information from overlapping panels of a repeated panel household survey. The coverage of the population by the individual panels at any given wave, and the use of the combined panels supplemented by a "top-up" sample to construct a representative cross-sectional sample are discussed in section 2. Also discussed in the same section are analogies with a multiple-frame survey scheme, as well as issues related to the sample dynamics. The weighting and estimation problem in repeated panel household surveys is described in section 3. Weighting strategies suitable for various panel survey situations are then proposed. Bias and efficiency issues related to the combination of panels are discussed. A weight adjustment procedure that deals effectively with changes in the combined panels over time is described in section 4. The integration of the various weight adjustments required for cross-sectional estimation in a repeated panel household survey is discussed in section 5. Finally, a summary and concluding remarks are provided in section 6.

[1] Takis Merkouris, Statistics Canada, Household Survey Methods Division, Tunney's Pasture, Ottawa, Ontario, K1A 0T6.

2. GENERAL CONSIDERATIONS

2.1 Coverage of the Cross-sectional Population

Important to cross-sectional estimation are changes in the population composition over time, occurring when individuals leave or enter the population. In a single-panel household survey, new entrants who have joined the survey population since the start of the panel are not represented in the sample at later waves if they live in households that do not contain any members of the original population. A multiple-panel household survey with overlapping panels provides a better coverage of the survey population than a single-panel survey, as it reduces the time period not covered by any of the panels. In the case of the SLID, this time period is reduced from a maximum of six years to a maximum of three years. Nevertheless, the problem of complete coverage remains unless a special supplementary sample of the non-covered population is taken at each survey wave. A survey scheme involving one panel and a supplementary sample drawn at each survey wave for cross-sectional purposes is described in Lavallée (1995). An alternative approach involves the selection, at each wave, of a new sample that covers the entire survey population but does not form a new panel. This sample (henceforth to be called top-up sample) is to be used only once, for cross-sectional purposes, and its size would normally be smaller than a panel's size. In the context of constructing a cross-sectional sample, a top-up sample is discussed as a non-trivial case of supplementary sample, essentially treated as an additional small overlapping panel.

The situation with regard to individuals who leave the population is as follows. For any panel, the sampling frame for the survey population at a time point t is essentially the sampling frame for the population at the start of the panel, with the leavers in the intervening period being treated as blanks on the frame. Panel members who leave the population before time t correspond to blanks on the frame, and thus their effect on cross-sectional estimates at time t is loss of efficiency but not bias; see also Kalton and Brick (1995) for relevant discussion.

The foregoing observations lead to the following perspective regarding the coverage of the population by each of the panels at any wave of the survey. As regards cross-sectional representation, each panel covers at the time of its selection the entire survey population represented by the preceding panels. Accordingly, the frames of the panels form a time sequence, with the frame of each panel containing at the start of the panel the frames of the preceding panels. In such a sequence of frames, a common frame is formed sequentially as the intersection of the frame of a new panel with the remainder of the original common frame of the preceding active panels. At any wave the common frame is the common frame at the start of the most recent of these panels, but without the leavers. The non-overlap frame domain at the start of a new panel consists of individuals who entered the population after the start of the preceding panel. Other frame domains (relatively very small in size) may be formed by returning units of older frames, in which case the time sequence of frames is not completely nested. Because of the latter type of frame domains, the complete frame at any wave after the selection of the most recent panel is the union of the frames of all panels at that time point, not just the remainder of the frame of the most recent panel. In panel surveys which employ a top-up sample at each wave the complete frame is that of the top-up sample.

2.2 A Multiple Frame Analogy

With the above considerations, a multiple panel survey with overlapping panels can be thought of as a special type of multiple frame survey, in which the frame for the cross-sectional population is the union of mutually exclusive temporal domains defined by the frames of the panels and their intersections. The sizes of the frames of the individual panels, as well as the characteristics of the population members in each panel's frame, change over time. This is in contrast with the static character of the usual type of multiple frame survey. Also, there is a high degree of nesting in the sequence of panel frames, so that the total number of mutually exclusive temporal frame domains is small. Among the various frame domains the one that is common to all panels is by far the largest. These special multiple frame features have implications in cross-sectional estimation, as will be discussed in the next section.

The sample temporal domains may be even less static because of attrition, moves of selected individuals within and between panels and moves of non-selected individuals into households in which panel members reside. For instance, with the presence of new entrants (*e.g.*, immigrants) in households that contain selected individuals, a panel crosses the boundary of its frame into the frame of the succeeding panel.

The analogy with multiple-frame survey sampling places the problem of cross-sectional estimation for repeated surveys with overlapping panels into a familiar framework. However, the distinctive dynamic features of multiple panel surveys will have to be considered if conventional multiple frame approaches are contemplated for the formulation of a cross-sectional estimation methodology.

For the purpose of introducing a cross-sectional estimation procedure that combines information from the panels of a repeated panel household survey, it suffices to consider the simple situation involving two overlapping panels at the time point of the start of the second panel. Note that this would always be the situation in a survey with one panel and a top-up sample. Thus, adopting standard multiple frame notation, with B and A denoting the frames of the first and the second panel ($B \subset A$) at the start of the second panel, and with s_B, s_A denoting the respective samples, the setting can be presented schematically as in Figure 1.

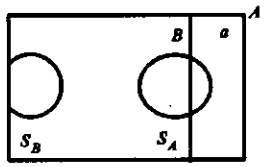

Figure 1. Two overlapping panels at the start of the second panel.

In Figure 1, A is the complete frame, so that the second panel at its start represents the cross-sectional population at that time. The overlap domain B is the remainder of the original frame of the first panel. The domain $a = B^c \cap A$ consists of all new entrants into the population since the start of the first panel. The samples s_B and s_A are the originally selected ones, with s_B reduced in size because of leavers and non respondents. It is assumed that the samples s_A and s_B are drawn independently from A and B according to specified probability designs $p_A(s_A)$ and $p_B(s_B)$, which determine the inclusion probabilities π_{Ai} and π_{Bi} of the i-th unit (household or any individual within it) for the original samples s_A and s_B, respectively. The samples s_A and s_B may intersect, since members in the overlap frame B can be selected in both panels. The issue of panel (sample) overlap is akin to that of duplicate sample units in multiple frame surveys. In repeated panel household surveys an operational constraint motivated by respondent burden may be to exclude from s_A individuals already selected in s_B, thus inducing $s_A \cap s_B = \varnothing$; for a discussion on this see Lavallée (1994). Here, as in any multiple frame situation, it is observed that if the probabilities π_{Ai} and π_{Bi} are small the probability of duplicate units is negligible. It will be assumed in the following that the probabilities π_{Ai} and π_{Bi} are small, and in effect $s_A \cap s_B = \varnothing$.

3. CROSS-SECTIONAL WEIGHTING AND ESTIMATION

This section describes procedures that combine information from multiple panels of a repeated panel household survey for cross-sectional estimation of population parameters. The discussion is confined to estimation of totals. A uniform approach to cross-sectional estimation for households and individuals is presented. This approach is based on the production of a set of weights for the combined panel sample that yield design-unbiased estimators of cross-sectional totals. Essentially, it involves the construction of a combined cross-sectional sample by means of an adjustment of the sampling weights of units from the temporal domains of the different panels that represent identical temporal domains of the cross-sectional population. While the delineation of the various temporal frame domains is necessary for determining the coverage of parts of the cross-sectional population by the different panels, the identification of some of the corresponding sample domains may not be possible under the operating procedures of a repeated panel household survey. For example, the information needed to determine whether or not a unit in the second panel belongs to the non-overlap frame domain a (see Figure 1) may not be available. In this section, both cases of identifiable and non-identifiable temporal sample domains are considered. The weight adjustment for the combination of the panels involves only sampled units, and takes no account of any changes (other than leavers) in household membership between waves. A "weight share" adjustment that handles such changes should follow the combination of the panels, as it can be applied readily only to the combined sample; see relevant discussion in section 4.

3.1 Identifiable Temporal Sample Domains

Weighting options for the combination of the panels

For the construction of a cross-sectionally representative combined sample, a panel survey scheme such as that depicted in Figure 1 is considered. In analogy with a standard multiple frame argument (Bankier 1986; Skinner and Rao 1996) the two samples s_A and s_B can be thought of as selected independently from the complete frame A according to the sampling designs $p_A(s_A)$ and $p_B(s_B)$, but with a fixed time lag between the two selections. Then the two sampling designs $p_A(s_A)$ and $p_B(s_B)$ induce a well-defined design $p(s)$ on the set of samples $s = s_A \cup s_B$ in A. Thus conventional estimators, based on a single frame and a combined sample, may be constructed from $p(s)$. The standard approach, leading to the Horvitz-Thompson estimator, would be to assign sample units weights made inversely proportional to their inclusion probabilities. The probability of inclusion of the i-th population unit in the combined sample, $\pi_i = P(i \in s)$, is equal to $\pi_{Ai} + \pi_{Bi} - \pi_{Ai}\pi_{Bi}$ if $i \in B$, and equal to π_{Ai} if $i \in a$. The weight of the i-th unit of the sample is then $w_i = 1/\pi_i$. This weighting scheme can be used provided that it is possible to identify the common units in the samples s_A and s_B, so that the duplicate units can be eliminated. A simpler approach, especially for surveys with more than two panels, would be to assign any unit $i \in B$ a weight made inversely proportional to the expected number of selections of the unit, that is, inversely proportional to $\pi_{Ai} + \pi_{Bi}$. This weighting scheme, proposed by Kalton and Anderson (1986) for multiple frame surveys, does not require identification of duplicate sample units. Now, consider the sample domains $s_{ab} = s_A \cap B$ and $s_a = s_A \cap a$ of s_A. Also, let a value y_i be associated with population unit i for some population characteristic, and define the population total $Y_A = \sum_A y_i \,(= \sum_B y_i + \sum_a y_i)$. Then, employing the latter weighting scheme the unbiased estimator

$$\hat{Y}_A = \sum_s w_i y_i = \sum_{s_B} (\pi_{Ai} + \pi_{Bi})^{-1} y_i + \sum_{s_{ab}} (\pi_{Ai} + \pi_{Bi})^{-1} y_i + \sum_{s_a} \pi_{Ai}^{-1} y_i \quad (1)$$

of the total Y_A can be constructed. On the assumption that the probabilities π_{Ai} and π_{Bi} for $i \in s \cap B$ are small, the estimator $\hat{Y}_A$ is approximately equal to the Horvitz-Thompson estimator.

The approach leading to the estimator (1) is not in general feasible, since the determination of the weight $w_i = (\pi_{Ai} + \pi_{Bi})^{-1}$ for $i \in s \cap B$ requires knowledge of π_{Ai} for units in s_B, and knowledge of π_{Bi} for units in s_{ab}. This is difficult or impossible to ascertain in household surveys because of stratified multistage sampling. In multiple-panel household surveys additional complications arise from the time element. For units that move (*e.g.*, to another stratum) in the time between the selection of the panels it is impossible to determine both π_{Ai} and π_{Bi}.

An alternative strategy needs to be considered for developing weights for the sample overlap domain $s \cap B$. One approach that provides a general framework for handling this problem requires information on the probability of inclusion in only one of s_A or s_B, thus avoiding the difficulty noted above. The essence of the alternative approach considered here is to associate with the i-th unit from the overlap frame B a number p_i $(0 \le p_i \le 1)$ when the unit is selected in s_B, and the number $1 - p_i$ when the unit is selected in s_A, and then define the weight of the unit as

$$w_i^* = p_i \frac{1}{\pi_{Bi}} I\{i \in s_B\} + (1 - p_i) \frac{1}{\pi_{Ai}} I\{i \in s_{ab}\}, \quad i \in B, \quad (2)$$

where I is the usual sample membership indicator variable. Clearly, $E(w_i^*) = 1$ under $p(s)$, and thus the use of the weights w_i^* will yield unbiased estimators $\hat{Y}_B = \sum_B w_i^* y_i$ for the total $Y_B = \sum_B y_i$, for any choice of constants p_i satisfying $0 \le p_i \le 1$, and for any sampling designs $p_A(s_A)$ and $p_B(s_B)$. Equation (2) can be written alternatively as $w_i^* = p_i w_{Bi} + (1 - p_i) w_{Ai}$, with the obvious definition of the weights w_{Bi} and w_{Ai} associated with the samples s_B and s_A. Thus, the class of weighting schemes defined by equation (2) consists essentially of different weighted combinations of the weights in the original samples s_B and s_A. The limits on the values of p_i ensure that the weight w_i^* will be nonnegative. Note that the intractable weight $w_i = (\pi_{Ai} + \pi_{Bi})^{-1}$, for $i \in s \cap B$, used in (1) is a special case of w_i^* with $p_i = \pi_{Bi}(\pi_{Ai} + \pi_{Bi})^{-1}$.

Evidently, the weighting scheme defined by (2) does not eliminate duplicate units that fall in both samples. If the operational constraint to exclude from s_A individuals already selected in s_B is imposed, the second term in the right-hand side of (2) should be modified to $(1 - p_i)$ $[\pi_{Ai}(1 - \pi_{Bi})]^{-1} I\{i \in s_{ab}, i \notin s_B\}$ to ensure that $E(w_i^*) = 1$. This, however, may be impossible to do since it requires that the inclusion probabilities of the sampled units be known over both frames. Note also that under the constraint of excluding duplicate units, the two samples will not be independent. Nevertheless, as it is assumed that both probabilities π_{Ai} and π_{Bi} are small, the probability of duplicate units will be negligibly small, and hence any bias resulting from using the tractable weighting scheme defined by (2) would also be negligible. On this assumption, the two indicator variables in (2) should be understood to satisfy $I\{i \in s_B\} I\{i \in s_{ab}\} = 0$.

The question arises now as to an optimal choice of p_i, for any $i \in s \cap B$, according to some criterion of optimal weighting for the combined sample. One approach is to choose the p_i to minimize the variance of the estimated total $\hat{Y}_A = \sum_B w_i^* y_i + \sum_a w_i y_i$, where $w_i = (\pi_{Ai})^{-1} I\{i \in s_a\}$ for $i \in a$. However, minimization of the variance of $\hat{Y}_A$ with respect to p_i for all $i \in s \cap B$ is not tractable. A simpler option is to restrict the class of weighting schemes defined by equation (2) to one in which the weight adjustment factors are specified not at the unit level but rather at a higher level, which may be a stratum or the entire overlap frame B. Further discussion on the level of adjustment is deferred to the last part of this subsection. It suffices for the development of the weighting procedure to consider next the case involving a uniform weight adjustment factor p for the entire frame B.

Determination of the value of p. Issues of practicality and efficiency.

The class of weighting schemes defined by equation (2) for the frame B, with uniform weight adjustment factor p, generates a class of unbiased estimators for the overall total Y_A of the form

$$\hat{Y}_A^p = p\hat{Y}_{s_B} + (1 - p)\hat{Y}_{s_{ab}} + \hat{Y}_{s_a}, \quad (3)$$

where $\hat{Y}_{s_B}$ and $\hat{Y}_{s_{ab}}$ are independent Horvitz-Thompson estimators of Y_B based on s_B and s_{ab}, respectively, and $\hat{Y}_{s_a}$ is the Horvitz-Thompson estimator of Y_a based on s_a. The limit values of p yield two special cases of the estimator $\hat{Y}_A^p$, in both of which the overlap domain total Y_B is estimated from one panel only. When p is set equal to zero in (3), the resultant trivial estimator $\hat{Y}_A^p$ for the entire population is based only on s_A. More notable is the case with p set equal to one in (3). The implied simple unbiased estimator $\hat{Y}_A = \hat{Y}_{s_B} + \hat{Y}_{s_a}$ would be the natural estimator in a panel survey with one panel and a supplementary cross-sectional sample, with the units in that sample being "screened" and only the units in the domain of new entrants being enumerated. In such a context this simple estimator would be a special case of a "screening" multiple frame estimator, the special feature being the temporal nature of the non-overlap frame domain a. In the present context the screening estimator appears inefficient because information in the sample domain s_{ab} is not utilized. Better use can be made of data from both panels by combining s_B and s_{ab}, using an optimal p that is based on the minimization of the variance of $\hat{Y}_A^p$. The optimal value of p is given by

$$p = \frac{\mathrm{Var}(\hat{Y}_{s_{ab}}) + \mathrm{Cov}(\hat{Y}_{s_{ab}}, \hat{Y}_{s_a})}{\mathrm{Var}(\hat{Y}_{s_B}) + \mathrm{Var}(\hat{Y}_{s_{ab}})}. \quad (4)$$

The variance and covariance terms in (4) are unknown, but could be estimated from the sample data, in which case the chosen p would actually minimize the estimated variance of $\hat{Y}_A^p$. There are many drawbacks associated with this choice of value for p. Generally, estimation of the optimal p is not easy; in surveys with more than two panels it would be very inconvenient to estimate the required set of such weight adjustments. Also, a sample estimate of the optimal p in (4) adds variability to the estimator $\hat{Y}_A^p$, and complicates the estimation of its variance. Moreover, the dependency of the estimated optimal p on the sample data entails $E(w_i^*) \neq 1$ for $i \in B$, which disturbs the unbiasedness of the estimator (3). It is to be noted that the condition $E(w_i^*) = 1$ is also necessary for the validity of the weight share method (see section 4) to hold when applied to the combined sample s at any wave after the selection of the second panel.

An alternative choice for the value of p is based on the minimization of the variance of the common-frame component $\hat{Y}_B^p = p\hat{Y}_{s_B} + (1-p)\hat{Y}_{s_{ab}}$ of the estimator $\hat{Y}_A^p$ in (3). This restricted minimization, which ignores the typically small domain estimator $\hat{Y}_{s_a}$, gives the value

$$p' = \frac{\mathrm{Var}(\hat{Y}_{s_{ab}})}{\mathrm{Var}(\hat{Y}_{s_B}) + \mathrm{Var}(\hat{Y}_{s_{ab}})}, \qquad (5)$$

which is independent of the covariance term, and always lies between zero and one. Minimizing the variance of $\hat{Y}_B^p$ conditional on the realized value of the random size n_{ab} of the sample domain s_{ab}, then using the well-known variance formula for the estimator of a total under simple random sampling, and disregarding finite population corrections, it can be shown that (5) may be approximated by

$$\hat{p}' = \frac{n_B/d_B}{n_B/d_B + n_{ab}/d_{ab}}, \qquad (6)$$

where n_B is the size of the sample s_B, and d_B, d_{ab} are the design effects associated with s_B and s_{ab}. The calculation of the value of $\hat{p}'$ requires estimates of the two design effects, which need not be based on s_B and s_{ab}. Suitable approximate values of d_B and d_{ab} may be available from other surveys with the same sampling designs as the two panels. However, because of the dependency of $\hat{p}'$ on the characteristic y through d_B and d_{ab}, a different set of weights needs to be calculated for each characteristic of interest. Besides making the estimation process operationally inconvenient, the different sets of weights may lead to inconsistencies among estimates. A compromise solution is to obtain approximate values of d_B and d_{ab} preferably for a count variable associated with a large population domain. A similar compromise solution is implicit in the approach of Skinner and Rao (1996) to estimation in dual frames. It is to be noted that since $\hat{p}'$ depends on the characteristic y only through the ratio d_B/d_{ab}, the loss of efficiency for estimators of totals of other characteristics should not be substantial. It is to be noted further that because of the time lag between the selection of the two panels, the design effects will be different, and thus present in (6), even when the sampling designs for the two panels are identical. By using estimates of the design effects from external sources the randomness of $\hat{p}'$ is due only to the random size of the sample domain s_{ab}. Since the size of the sample s_A is usually very large, and the size of the overlap frame B is typically only a little smaller than the size of the complete frame A, the size n_{ab} of the sample domain s_{ab} must be nearly constant, and thus the unbiasedness condition $E(w_i^*) = 1$ will hold approximately.

Some loss of efficiency will be incurred by ignoring $\hat{Y}_{s_a}$ in deriving an optimal value for p, but this loss may be insignificant given the relatively very small size of the domain a in most household panel surveys, because of the typically small time lag between panels. To assess this loss of efficiency, let $\hat{Y}_A^p$ and $\hat{Y}_A^{p'}$ denote the estimator $\hat{Y}_A^p$ in (3) with the value of p given by (4) and (5), respectively. Then, a simple calculation gives

$$\begin{aligned}\mathrm{Var}(\hat{Y}_A^{p'}) - \mathrm{Var}(\hat{Y}_A^p) &= \frac{\mathrm{Cov}^2(\hat{Y}_{s_{ab}}, \hat{Y}_{s_a})}{\mathrm{Var}(\hat{Y}_{s_B}) + \mathrm{Var}(\hat{Y}_{s_{ab}})} \\ &\leq \frac{\mathrm{Var}(\hat{Y}_{s_{ab}})\mathrm{Var}(\hat{Y}_{s_a})}{\mathrm{Var}(\hat{Y}_{s_B}) + \mathrm{Var}(\hat{Y}_{s_{ab}})} \\ &= p'\,\mathrm{Var}(\hat{Y}_{s_a}),\end{aligned}$$

so that an upper bound for the efficiency loss can be obtained as

$$\frac{\mathrm{Var}(\hat{Y}_A^{p'}) - \mathrm{Var}(\hat{Y}_A^p)}{\mathrm{Var}(\hat{Y}_A^p)} \leq p' \frac{\mathrm{Var}(\hat{Y}_{s_a})}{\mathrm{Var}(\hat{Y}_A^p)}.$$

Given the usually very small size of $\hat{Y}_{s_a}$ relative to $\hat{Y}_A^p$ (the size of the domain a is approximately one fortieth of the size of the complete frame A in the case of the SLID) it appears that the loss of efficiency will be very small in most panel household surveys.

An interesting question is whether or not $\hat{Y}_A^{p'}$ is more efficient than the simple "screening" estimator $\hat{Y}_A = \hat{Y}_{s_B} + \hat{Y}_{s_a}$, whose variance is $\mathrm{Var}(\hat{Y}_{s_B}) + \mathrm{Var}(\hat{Y}_{s_a})$. It can be readily shown that $\mathrm{Var}(\hat{Y}_A^{p'}) < \mathrm{Var}(\hat{Y}_{s_B}) + \mathrm{Var}(\hat{Y}_{s_a})$ if $2\mathrm{Cov}(\hat{Y}_{s_{ab}}, \hat{Y}_{s_a}) < \mathrm{Var}(\hat{Y}_{s_B})$. This condition certainly holds if the covariance of $\hat{Y}_{s_{ab}}$ and $\hat{Y}_{s_a}$ is negative, which may be the case if the estimated characteristic differs between immigrants versus non immigrants. In general, this covariance may actually be positive because $\hat{Y}_{s_{ab}}$ and $\hat{Y}_{s_a}$ are based on the same sampled area clusters. In that case too, however, the condition will most likely hold, given the magnitude of $\mathrm{Var}(\hat{Y}_{s_B})$ relative to $\mathrm{Var}(\hat{Y}_{s_{ab}})$, and the magnitude of $\mathrm{Var}(\hat{Y}_{s_{ab}})$ relative to $\mathrm{Var}(\hat{Y}_{s_a})$. Indeed, the

sizes of the panel samples s_B and s_A are typically equal by design, although the effective panel sizes (*i.e.*, realized sizes at any wave, adjusted for design effects) may be considerably different due to different attrition rates and design effects for the two panels. Also, with the sizes of the sample domains s_{ab} and s_a roughly proportional to the corresponding population domain sizes, $\mathrm{Var}(\hat{Y}_{s_a})$ will be many times, say k, smaller than $\mathrm{Var}(\hat{Y}_{s_{ab}})$. Then,

$$2\,\mathrm{Cov}(\hat{Y}_{s_{ab}}, \hat{Y}_{s_a}) \le 2\sqrt{\mathrm{Var}(\hat{Y}_{s_{ab}})\mathrm{Var}(\hat{Y}_{s_a})}$$

$$= 2\,\frac{\mathrm{Var}(\hat{Y}_{s_{ab}})}{\sqrt{k}},$$

so that a sufficient condition for the estimator $\hat{Y}_A^{p'}$ to be more efficient than the "screening" estimator is

$$2\,\frac{\mathrm{Var}(\hat{Y}_{s_{ab}})}{\sqrt{k}} < \mathrm{Var}(\hat{Y}_{s_B}).$$

The interpretation of this is that the sample domain s_{ab} is not to be ignored when estimating Y_A if $\mathrm{Var}(\hat{Y}_{s_B})$ is not too small relative to $\mathrm{Var}(\hat{Y}_{s_{ab}})$. The condition is ordinarily satisfied in panel household surveys. An additional argument in favour of including s_{ab} in estimation is its better quality relative to s_B, since the latter is more liable to the potential bias effect of sample attrition.

The simple approximate weight adjustment factor $\hat{p}'$ given by expression (6) affords an efficient combination of panel samples, accounting for the precision of $\hat{Y}_{s_B}$ relative to that of $\hat{Y}_{s_{ab}}$ through the effective sample sizes n_B/d_B and n_{ab}/d_{ab}. These effective sample sizes are time-dependent, though their ratio (and hence $\hat{p}'$) should be quite stable over the period of panel overlap. Regarding variance calculations, since n_{ab} is typically nearly non-random, the adjustment factor $\hat{p}'$ can be conveniently treated as constant in any variance estimation procedure.

It is important to emphasize here that additional gains in efficiency will result from the incorporation of auxiliary information into the weights through a calibration weight adjustment to known population totals.

Finally, it should be remarked that if the criterion in the choice of the value of p is the minimization of the mean square error of the common-frame component $\hat{Y}_B^p = p\hat{Y}_{s_B} + (1-p)\hat{Y}_{s_{ab}}$ of the estimator $\hat{Y}_A^p$, then it can be easily shown that when the biases of $\hat{Y}_{s_B}$ and $\hat{Y}_{s_{ab}}$ are equal the optimal value of p is the same as the one given by (5). The biases are not expected to be equal, though; for instance, the different sample attrition rates for the two panels may result in different levels of bias. It is clear that the bias of the linear combination $\hat{Y}_B^p = p\hat{Y}_{s_B} + (1-p)\hat{Y}_{s_{ab}}$, though not minimized if p is as in (5), is nevertheless smaller than the larger of the two component biases. Other complexities aside, the unavailability of good estimates for the two biases renders the criterion of minimum mean square error impracticable.

Generalization to multiple panels and discussion of alternative approaches.

The weighting procedure described above applies to the simple situation of a two-panel survey at the start of the second panel. At later survey waves an additional non-overlap frame domain, denoted by b, may be formed by returning leavers of the frame B. Units from b originally selected in the first panel were not present when the second panel was selected. Clearly, the weights in the non-overlap sample domain s_b are not to be adjusted for the purpose of combining the two panels. Furthermore, the value for p will not be affected, as it is based only on the overlap domain of the combined sample. As with ignoring the sample domain s_a in determining the value of p, ignoring the much smaller, possibly void, sample domain s_b will have negligible impact on the efficiency of derived estimators.

The simplicity of the proposed weighting procedure for the combination of two panels makes its generalization to surveys with more than two overlapping panels straightforward. The most likely generalization in practice would involve three panels. The construction of a combined cross-sectional sample would then involve the adjustment of the sampling weights of units from temporal domains of the different panels that represent a common temporal domain of the cross-sectional population. For each common temporal population domain the weight adjustment factors will be based on the relative effective sample sizes of the corresponding panel domains, in analogy with expression (6), and will add up to one. The number of common temporal frame domains, and hence the number of the corresponding independent sets of adjustment factors, will be quite small because of the high degree of nesting in the sequence of panel frames. For instance, for a three-panel survey there will be one set of three adjustment factors and one set of two.

Returning now to an earlier point, varied weight adjustment factors may be specified at a lower level of sample grouping, such as a certain stratification level. For reasons of feasibility (identical stratification for the two panels is required for that level) and operational convenience, a high level of stratification should be chosen. The natural choice is a superstratum level, at which all other weighting and estimation procedures are carried out independently for each superstratum. In the SLID, such superstrata are the Canadian provinces. The advantage of specifying weight adjustment factors at the superstratum level is improved efficiency, since an optimal or nearly optimal weight adjustment factor p can be determined for each superstratum. This will be particularly advantageous if the ratios of the effective sample sizes of the panels are very different among the superstrata, as is the case in the SLID.

Alternative estimation techniques from the general theory of multiple frame surveys with complex designs (for an account, see Skinner and Rao 1996, and Singh and Wu 1996) would produce estimators similar in form to the

estimator (3) if adapted to a multiple panel survey with overlapping panels. Such techniques, though, are not preferable in general for reasons similar to those stated in the discussion following equation (4); the "pseudo-likelihood" method of Skinner and Rao (1996) is also not applicable in surveys with more than two panels. Furthermore, while the weight adjustment proposed in this section essentially combines the panels, on the basis of an efficient combination of Horvitz-Thompson estimators, the standard multiple frame methods ordinarily combine ratio-adjusted or, more generally, calibrated estimators derived separately using the sample from each frame. In the context of a household panel survey, the components from each panel would be calibrated estimators incorporating all the weight adjustments, including the "weight share" adjustment, carried out separately for each panel. This would be in conflict with the application of the "weight share" adjustment to the combined sample, to be proposed in section 4. It is interesting to note that apart from this complication there are many possible limitations that could render a separate calibration of each panel problematic or unfeasible. It may be remarked first that a proper separate calibration of the panels is possible only when the various temporal sample domains are identifiable. Furthermore, a calibration involving the same auxiliary variables for each temporal domain of each panel would be required in order for the final weights to satisfy all calibration constraints. But since all temporal frame domains (except the one that is common to all panels) are typically very small, a calibration involving a large number of auxiliary totals (as is customary in household surveys) would not be sensible for reasons of potential bias and loss of efficiency of derived estimators. Moreover, auxiliary totals for frames of old panels that account for the loss of population units may not be available. It should also be pointed out that accurate auxiliary totals most likely would be unavailable if the frame of each panel were augmented with new entrants who live with individuals of the original frame of the panel. Such would be the situation if the "weight share" procedure, which assigns a basic weight to new entrants living with selected individuals, were to precede the combination of the panels.

Notwithstanding other difficulties, it is possible in principle to use standard multiple frame methods to combine the panels, avoiding a separate calibrating weight adjustment, with the exception of the dual-frame pseudo-likelihood method of Skinner and Rao which in the setting of Figure 1 would require a simple ratio weight adjustment for s_B, s_{ab} and s_a.

Lastly, a known drawback of various multiple frame estimators is that their optimality depends on the estimated characteristic of interest. For the proposed method this dependency appears to be weaker, because the optimal $\hat{p}'$ in (6) depends on the particular characteristic only through a ratio of panel design effects, estimated from an extraneous source.

3.2 Non-identifiable Temporal Sample Domains

It has been assumed thus far that the units of the non-overlap sample domain $s_a (\subset s_A)$ can be identified. However, the information needed to determine whether a unit in s_A belongs to the frame domain a, of new entrants into the population after the start of the previous panel, may not be available for all units of s_A. In that situation the weighting process described above would combine the two samples s_B and s_A without distinguishing between the domains s_{ab} and s_a of s_A, so that the weights of units in s_a would also be multiplied by $1 - p$. The estimator $\hat{Y}_A^p$ in (3) would collapse then to

$$\hat{Y}_A^p = p\hat{Y}_{s_B} + (1-p)\hat{Y}_{s_A}. \tag{7}$$

The effect of this error is the underestimation of the total Y_a for the population domain a by the factor p. Part of the domain a, though, consists of newborns, which can be identified in s_A with certainty. Their weights could very well be excluded from the adjustment by the factor $1 - p$, but that would have no effect on cross-sectional estimation, unless newborns were part of the population of interest. Besides, adjusting the weights of newborns in s_a by the factor $1 - p$ has the desirable effect of producing a common household weight. A calibration of the weights of the combined sample to known population totals of the complete frame A will lessen the under-representation of the rest of the domain a, which consists mainly of immigrants, but some bias may still result if the survey characteristics of the members of this part of the population are quite different from those of the members of the population domain B. Unless the time lag between the selection of the two panels is quite large, the size of this part of the population is very small, relative to the total population, and the potential bias effect on overall estimates of totals should be negligible.

The optimal (*i.e.*, variance minimizing) value of p in (7) is given now by

$$p'' = \frac{\mathrm{Var}(\hat{Y}_{s_A})}{\mathrm{Var}(\hat{Y}_{s_A}) + \mathrm{Var}(\hat{Y}_{s_B})}. \tag{8}$$

Disregarding finite population corrections it can be shown that (8) can be expressed as

$$\hat{p}_c'' = \frac{n_B d_A N_A^2 S_A^2}{n_B d_A N_A^2 S_A^2 + n_A d_B N_B^2 S_B^2} = \frac{n_B d_A}{n_B d_A + c\, n_A d_B}, \tag{9}$$

with $c = (N_B^2 S_B^2)(N_A^2 S_A^2)^{-1}$, and where n_B, n_A are the sizes of the samples s_B and s_A; d_B, d_A are the design effects associated with s_B and s_A and the characteristic y; N_A, N_B are the sizes of the frames A and B; S_A^2, S_B^2 are the variances of the characteristic y in A and B. Noting that N_B may be only a little smaller than N_A (depending on the time lag between the two panels), and assuming that the unknown variances S_A^2 and S_B^2 are nearly equal, a good practical approximation of the optimal p can be obtained by simply setting c equal to one in (9). The assumption that the variances S_A^2 and S_B^2 are nearly equal is reasonable considering the magnitude of N_B relative to that of N_A. Approximate values of d_B and d_A available from other surveys with the same designs as the two panels could be used, preferably for a characteristic such as the size of a large population domain. Now, if $\hat{Y}_c$ and $\hat{Y}_1$ denote the estimator $\hat{Y}_A^p$ in (7) when the weight adjustment $\hat{p}_c''$ in (9) is used with the true value of c and the approximate value $c = 1$, respectively, then ignoring finite population corrections the loss of efficiency of $\hat{Y}_1$ relative to $\hat{Y}_c$ can be readily shown to be

$$\frac{\mathrm{Var}(\hat{Y}_c) - \mathrm{Var}(\hat{Y}_1)}{\mathrm{Var}(\hat{Y}_c)} = -\frac{(c-1)^2}{c}\hat{p}_1''(1-\hat{p}_1'').$$

With a value of c most likely in the neighbourhood of 1.0, the loss of efficiency will be negligible.

It is interesting to examine the efficiency of the estimator given by (7), with p'' as in (8), relative to the optimal estimator given by (3), with p as in (4), used when the domain s_a is identifiable. Let $\hat{Y}_A''$ and $\hat{Y}_A$ denote these estimators, respectively. Then, using the inequality $\mathrm{Cov}^2(\hat{Y}_{s_A}, \hat{Y}_{s_{ab}}) \le \mathrm{Var}(\hat{Y}_{s_A})\mathrm{Var}(\hat{Y}_{s_{ab}})$ it can be shown that $\mathrm{Var}(\hat{Y}_A) - \mathrm{Var}(\hat{Y}_A'') \ge (p'' - p')\mathrm{Var}(\hat{Y}_{s_A})$, where p' is as in (5). As already mentioned, in general $\mathrm{Cov}(\hat{Y}_{s_{ab}}, \hat{Y}_{s_a}) > 0$, so that $p'' > p'$ and hence $\mathrm{Var}(\hat{Y}_A) \ge \mathrm{Var}(\hat{Y}_A'')$. Therefore, notwithstanding the use of the exact values of p'' and p' in the comparison, the approach taken in this subsection may in most cases result in reduction of the variance of derived estimators. A lower bound for the gain in efficiency relative to $\hat{Y}_A$ would then be given by

$$\frac{\mathrm{Var}(\hat{Y}_A) - \mathrm{Var}(\hat{Y}_A'')}{\mathrm{Var}(\hat{Y}_A)} \ge \frac{(p'' - p')}{1 - p'}.$$

An extension of the weight adjustment procedure described above to surveys involving more than two panels with non-identifiable temporal sample domains is straightforward. There will be then as many weight adjustment factors, adding up to one, as there are panels. This very practical procedure will produce good cross-sectional estimates in multiple panel surveys in which the time lag between the selection of the panels is not large. Otherwise, the potential for bias due to the domain identification error may be of concern, mainly for estimates related to subpopulations composed in substantial proportion of new entrants.

4. THE WEIGHT SHARE METHOD FOR THE COMBINED PANELS

This section describes the application of a weight adjustment method, known as the weight share method, to the combined panel sample at any wave after the start of the most recent panel. This weight adjustment is necessary because of the changes in the household membership after the selection of the panels.

The weight share method is a cross-sectional weighting procedure that assigns a basic weight to every individual in a panel household at any wave after the first. In particular, the weight share method, as applied to a single panel, assigns a positive weight to non-selected individuals who join households containing at least one individual selected for the original sample. Following Lavallée (1995), in this paper such households are termed longitudinal households, while the non-selected individuals living in longitudinal households are termed cohabitants. The cohabitants are distinguished into originally present cohabitants if they belong to the original (sampled) population, and originally absent cohabitants if they are new entrants to the population. Other problematic situations that can be handled by the weight share method involve non-selected households formed after the first wave by members of separate originally selected households, as well as originally selected individuals who have subsequently moved to other longitudinal households. For a detailed discussion of the weight share method for a single panel, see Kalton and Brick (1995), and Lavallée (1995). For the purpose of applying the weight share method to a multiple panel survey the following need to be considered. In multiple panel surveys, the original population for the combined panels is the union of the populations covered by the different panels at the time of their selection. Accordingly, the original sample consists of all selected units in the combined panel sample. Thus, an originally present cohabitant is an individual that was eligible for selection in any of the panels. In this approach then, at any wave after the selection of the most recent panel a cohabitant is distinguished into originally present or originally absent with respect to the original combined panel sample, not with respect to each original panel. Notably, at the first wave of a new panel, or when a top-up sample is used, all cohabitants are originally present. On the other hand, application of the weight share method separately to each panel (before combination) would require more precise information on the eligibility of the cohabitants for selection in each of the various panels, in order to distinguish the originally present cohabitants from the originally absent cohabitants and to identify the temporal domain that includes each of the cohabitants. Such information most likely would be unavailable. Moreover, combining the panels after the weight share

procedure would require a very complicated set of specifications in order to ensure that a suitable weight adjustment factor would be applied to each sampled unit. For instance, with the inclusion of the originally absent cohabitants into the panels through the weight share procedure, the frames of the panels will be different at each survey wave, thereby complicating the determination of the various temporal domains. Lastly, it should also be pointed out that in multiple panel surveys sampled individuals may move from one panel to another panel between waves during the time period of panel overlap, and non-sampled households may be formed by members of originally selected households from different panels. Thus, the panels are truly distinct (and independent) only with respect to the time of their selection.

It follows from the foregoing considerations that the weight share method for multiple panels is to be applied to the combined panel sample, and not to each panel separately. Then, with the prescribed distinction of the two types of cohabitants, the case of the weight share method for a multiple panel survey reduces to the case of a single panel survey. As a desirable consequence, the application of the weight share method to the combined sample will yield always a common weight for all members of the same household. The following is an exemplification of the proposed weight share procedure for multiple panel surveys, involving the simple case of two panels.

Starting with a survey setting as depicted in Figure 1, with two overlapping panels at the time point of the start of the second panel, let there be N individuals in the population at a later wave (time t), with N_i individuals in household $\mathcal{H}_i$, say; $i = 1,...,H$ and $\sum N_i = N$. Let M_i denote the number of individuals in household $\mathcal{H}_i$ at time t that belong to the original population, with M_{Bi} and M_{ai} individuals from the original frame domain B and the non-overlap frame domain a, respectively, so that $M_i = M_{Bi} + M_{ai}$. Some, but not all, of the numbers M_{Bi}, M_{ai} and $N_i - M_i$ may be zero for any particular household. Now, with the random weights of individuals in B and a as defined in section 3.1, and with the weights of the $N_i - M_i$ originally absent cohabitants in $\mathcal{H}_i$ being identically equal to zero, the weight share method defines a common weight for every individual in $\mathcal{H}_i$ (including new members) as

$$w_i = \frac{1}{M_i} \sum_{k=1}^{M_i} w_{ik}, \qquad (10)$$

where w_{ik} is the weight of the k-th household member that belongs to the original population. Clearly then $E(w_i) = 1$ for each household for which $M_i \neq 0$, whereas $E(w_i) = 0$ if $M_i = 0$, since $w_i \neq 0$ only if $M_i > 0$. For the survey characteristic y, the total for the population of individuals at time t can be expressed as $Y = \sum_{i=1}^{H} \sum_{k=1}^{N_i} y_{ik}$, where y_{ik} is the value of y for individual k in household $\mathcal{H}_i$. Then, an estimator of Y is given by

$$\begin{aligned} \hat{Y} &= \sum_{i=1}^{H} w_i \sum_{k=1}^{N_i} y_{ik} \\ &= \sum_{i=1}^{H} w_i \left[\sum_{k=1}^{M_{Bi}} y_{ik} + \sum_{k=1}^{M_{ai}} y_{ik} + \sum_{k=1}^{N_i - M_i} y_{ik} \right] \\ &= \hat{Y}_B + \hat{Y}_a + \hat{Y}_{A^c}, \end{aligned} \qquad (11)$$

with w_i as in (10), with A^c denoting the set of individuals not in frame A, and with the obvious notation for the right hand side of (11). The estimator $\hat{Y}$ in (11) is given as the sum of three estimators, $\hat{Y}_B$, $\hat{Y}_a$ and $\hat{Y}_{A^c}$, for the totals related to the population domains B, a and A^c, respectively. The estimators $\hat{Y}_B$ and $\hat{Y}_a$ are unbiased, even though they are based on sets of units that may not be identical to the original samples $s_B \cup s_{ab}$ and s_a, respectively. For example, the estimator $\hat{Y}_B$ is based on a set of units consisting of the remaining units of the original combined sample $s_B \cup s_{ab}$ from frame B, and possibly of cohabitants originally present in B. The estimator $\hat{Y}_{A^c}$ is not unbiased for Y_{A^c}, because individuals in A^c who live in households that contain no members of the original population are not represented in the panel survey. Nevertheless, the estimator $\hat{Y}_{A^c}$ is unbiased for the total corresponding to the rest of A^c, which is represented in the combined panels by the originally absent cohabitants. In the special case when time t coincides with the start of the second panel (or with the time of selection of a supplementary sample), $A^c = \varnothing$, $N_i = M_i$, and the estimator $\hat{Y} = \hat{Y}_B + \hat{Y}_a$ is unbiased for Y. It should be noted here that if the weights of the responding individuals at time t are adjusted for nonresponse, the relationship $E(w_i) = 1$ may hold only approximately, and in that sense the resulting estimators may be only approximately unbiased.

It is important to note that the estimator $\hat{Y}$ in (11) can be expressed as

$$\hat{Y} = \sum_{i=1}^{H} w_i Y_i,$$

where $Y_i = \sum_{k=1}^{N_i} y_{ik}$ is the total for household $\mathcal{H}_i$. Thus, $\hat{Y}$ is also an estimator of the household-level total at time t.

As with the weight adjustment involved in the combination of panels, the weight share adjustment may also be carried out at a superstratum level, say province, for the combined sample of each province. In this approach, those individuals who at time t reside in a province other than the one in which they resided at the time of selection of any of the panels are treated as originally absent, since they were not members of the original population of their new province. In particular, interprovincial movers (selected or non selected in their original province) who are found in longitudinal households in their new province at time t are treated as originally absent cohabitants. When a top-up sample is used at time t, these interprovincial movers are

treated as originally present cohabitants. The application of the weight share procedure separately for each superstratum enjoys certain operational and statistical advantages over the standard weight share procedure. An account of the comparative merits of the two approaches is given in Merkouris (1999).

5. INTEGRATION OF VARIOUS WEIGHT ADJUSTMENTS

In addition to the weight adjustments described so far, other adjustments to the weights of a panel household survey may also be required. The integration of the various weight adjustments is briefly outlined below.

The first adjustment, applied in relation to the original sample units, is for wave nonresponse, which arises when a sampled unit responds for some but not all of the waves for which it was eligible. For a discussion on weight adjustment for wave nonresponse, see Kalton and Brick (1995). The adjustment is made separately to the different panels at each wave.

The second adjustment is for the combination of the samples of the various panels into one sample for cross-sectional estimation. It applies to the weights of the sampled units of the panels, adjusted for wave nonresponse, and employs the method described in section 3.

The third adjustment involves the application of the weight share procedure to the combined panel sample at any wave after the start of the most recent panel, as described in section 4.

Finally, in the weight calibration adjustment the weights of the combined panel units are adjusted so as to make the estimated totals for certain auxiliary characteristics equal to known population totals for these characteristics at the current wave, which in the simple case as in Figure 1 correspond to totals of the complete frame A. In more general situations, after the selection of the most recent panel the calibration totals will include the new entrants into the population. Note that in the absence of a top-up sample the new entrants will be represented in the panels only by the originally absent cohabitants. Calibrating the weights of the combined sample to population totals of each of the different temporal domains (when the panel units from these domains can be identified) may not be feasible or sensible for reasons already noted in section 3.1.

6. SUMMARY AND CONCLUDING REMARKS

The weighting procedures described in this paper can be used to combine information from multiple panels of a repeated household survey for cross-sectional estimation in a fairly general setting involving panels with given designs; design issues regarding determination of optimal sampling fractions for the panels, in conjunction with efficient combination of the panel data, are beyond the scope of this paper. It has been shown that although a multiple panel survey can be viewed as a special type of multiple frame survey, its distinctive dynamic character renders conventional multiple frame estimation procedures problematic or even non applicable. The proposed weighting procedures, which account for the population and panel dynamics, involve a simple weight adjustment for each panel that is proportional to the effective panel size. These procedures are operationally convenient for any number of overlapping panels, and for different situations regarding the identifiability of various temporal panel domains. Theoretical and practical issues related to the application of a weight share adjustment, to the calibration weight adjustment and to the integration of the various weighting procedures involved in a multiple panel survey have also been addressed. In particular, it has been argued that the weight adjustment for the combination of the panels should precede the weight share adjustment, with calibration being the final weight adjustment. A detailed empirical study of issues pertaining to the determination of weight adjustment factors for combining two panels of the SLID, based on the methodology of this paper, is described in Latouche *et al.* (2000). The variance of cross-sectional estimators has been discussed in this paper only in the context of efficient combination of panels. Variance estimation issues related to changes in the sample over time, particularly to moves from stratum to stratum, are discussed in Merkouris (1999). It is to be remarked, in conclusion, that the quality of a cross-sectional estimation procedure depends on the identifiability of various overlap temporal sample domains; on design features of the survey, such as the duration of (and the lag between) the panels and the use of a supplementary sample at any survey wave; and on the adequacy of the information on cohabitants required for the application of the weight share method.

ACKNOWLEDGEMENTS

The author wishes to thank Milorad Kovacevic, Michel Latouche, Pierre Lavallée and Harold Mantel for useful comments. Detailed comments and suggestions by three referees on an earlier version of this paper improved both its content and its presentation.

REFERENCES

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

DEVILLE, J. C. (1998). Les enquêtes par panel : En quoi diffèrent-elles des autres enquêtes? Suivi de comment attraper une population en se servant d'une autre. *Actes des Journées de méthodologie statistique*, numéro 84-85-86, 63-82.

KALTON, G., and ANDERSON, D. W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society*, A, 149, 65-82.

KALTON, G., and CITRO, C. F. (1993). Panel Surveys: Adding the fourth dimension. *Survey Methodology*, 19, 205-215.

KALTON, G., and BRICK, J. M. (1995). Weighting schemes for household panel surveys. *Survey Methodology*, 21, 33-44.

LATOUCHE, M., DUFOUR, J. and MERKOURIS, T. (2000). Cross-sectional weighting for the SLID: Combining two or more panels. Income Research Paper Series, 75F0002MIE6, Statistics Canada.

LAVALLÉE, P. (1994). Ajout du second panel à l'EDTR : sélection et pondération. Internal document, Statistics Canada.

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

LAVIGNE, M., and MICHAUD, S. (1998). General aspects of the Survey of Labour and Income Dynamics. Working Paper SLID 98-05 E, Statistics Canada.

MERKOURIS, T. (1999). On the weight share method for panel household surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 255-260.

SINGH, A.C., and WU, S. (1996). Estimation for multiframe complex surveys by modified regression. *Proceedings of the Survey Methods Section, Statistical Society of Canada*, 69-77.

SKINNER, C.J., and RAO, J.N.K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91, 349-356.

Survey Methodology, December 2001
Vol. 27, No. 2, pp. 183-188
Statistics Canada

Producing Small Area Estimates From National Surveys: Methods for Minimizing use of Indirect Estimators

DAVID A. MARKER[1]

ABSTRACT

National surveys are usually designed to produce estimates for the country as a whole and for major geographical regions. There is, however, a growing demand for small area estimates on the same attributes measured in these surveys. For example, many countries in transition are moving away from centralized decision-making, and western countries like the United States are devolving programs such as welfare from Federal to state responsibilities. Direct estimates for small areas from national surveys are frequently too unstable to be useful, resulting in the desire to find ways to improve estimates for small areas. While it is always possible to produce indirect, model-dependent, estimates for small areas, it is desirable to produce direct estimators where possible. Through stratification and oversampling, it is possible to increase the number of small areas for which accurate direct estimation is possible. When estimates are required for other small areas, it is possible to use forms of dual-frame estimation to combine the national survey with supplements in specific areas to produce direct estimates. This article reviews the methods that may be used to produce direct estimates for small areas.

KEY WORDS: Small area estimation; Direct estimation; Stratification; Oversampling; Dual-frame estimation.

1. INTRODUCTION

Throughout the world there is an increased demand for small area estimates. During the 1990s countries in transition moved away from centralized decision-making, requiring accurate estimates of local economic and demographic conditions. In the United States the Federal government has been moving responsibility for many social programs to the 50 states. Evaluating the success of such efforts requires accurate estimates for each state. Some programs such as the Small Area Income and Poverty Estimates (Citro and Kalton 2000) are required at much smaller levels of geography, for example for thousands of school districts. Regardless of the best plans of survey designers, "The client will always require more than is specified at the design stage" (Fuller 1999, page 344).

Ideally such estimates would be produced from direct (design-based) estimators. Unfortunately, at small levels of aggregation, the direct estimates are too unstable to be published and/or used for policy purposes. As a result there has been a great deal of interest in developing a range of indirect estimation techniques (Marker 1999; Rao 1999; Ghosh and Rao 1994).

This paper approaches this problem from a different perspective, how to minimize model-reliance through good survey design. It will never be possible to anticipate all survey uses, or to allocate sufficient sample sizes to all domains of interest, so indirect estimators will always be needed. It is possible, however, to make design choices that will greatly improve the ability of national surveys to support direct estimation for many small areas. Such choices can also improve the ability of surveys to be used to produce indirect estimates where they are needed. This paper is an update of the excellent paper by Singh, Gambino and Mantel (1994) on the same topic. Design issues that will be considered include stratification and oversampling, combining multiple years of data, harmonization across surveys, dual-frame estimation, and measuring the accuracy of estimates.

2. STRATIFICATION AND OVERSAMPLING

Deciding on the optimal stratification and oversampling scheme for any national survey is a compromise across many variables of interest. Optimizing stratification and oversampling between national estimates and small area estimates should also be a compromise. By giving up some national accuracy it is often possible to greatly improve the accuracy for many small areas. Some of these small areas may then be able to support accurate design-based estimates. Other small areas will still require model assistance, but the stratification may allow for unbiased (but variable) estimates that can be incorporated into the model-based estimates. As the following example demonstrates, stratification alone is helpful, but limited, in its ability to improve small area estimates.

The United States Current Population Survey (CPS), conducted by the U.S. Census Bureau, has stratified by state and unemployment rate since 1985. However, another large Census Bureau survey, the United States National Health Interview Survey (NHIS), stratified by region, metropolitan area status, labor force data, income, and racial composition until 1994. The resulting sample sizes for individual states varied from year to year and did not support unbiased state-level estimates. Due to random sampling, from 1985

[1] David A. Marker, Westat, 1650 Research Blvd., Maryland, U.S.A. 20850, e-mail: DavidMarker@Westat.com.

to 1994 two states did not have any sample included in the NHIS. This would not have happened with state stratification.

Beginning in 1995 the NHIS stratification scheme was replaced by state and metropolitan status. Table 1 summarizes the number of states that have sufficient sample size in the 1995 NHIS to achieve various levels of accuracy for four different key health measures. The NHIS completes interviews with approximately 44,000 households containing 100,000 individuals. With a very strict constraint of a 10 percent coefficient of variation (CV) less than 10 states meet the standard for three of the four variables. Over half of the states meet the more lenient 30 percent CV for all four variables, but even this standard is not met for all states.

Figure 1 presents the ability of the NHIS to meet these accuracy standards for generic questions with prevalence levels of 0.01, 0.05, 0.10, 0.15, and 0.20 and design effects ranging from 1.00 to 6.00. (This variation in design effects is found on the NHIS, depending on the intra-household correlation and other clustering.) For prevalence rates above 10 percent, almost all states can achieve the 30 percent criterion even for the largest design effects. However, there is a significant drop off in the number of states as the criterion is tightened, the design effect increases, or the prevalence rate drops. For rare events with even moderate design effects less than half the states can meet the weakest criterion and hardly any can make the tightest.

Table 1
Summary of the Number of States (out of 51, Including the District of Columbia) That Have the Required 1995 NHIS Sample Size to Achieve a CV of 30-, 20-, and 10-Percent for Four Selected Variables (44,000 Households, 100,000 Individuals)

Coefficient of Variation (CV)	Percent uninsured: all ages (p = 13.5%)	Percent uninsured: under 19 (p = 12.2%)	Percent uninsured: low income children (p = 20.4%)	Percent smokers: 18 and over (p = 25.2%)
30-percent	42	31	28	45
20-percent	31	13	10	36
10-percent	7	2	2	14

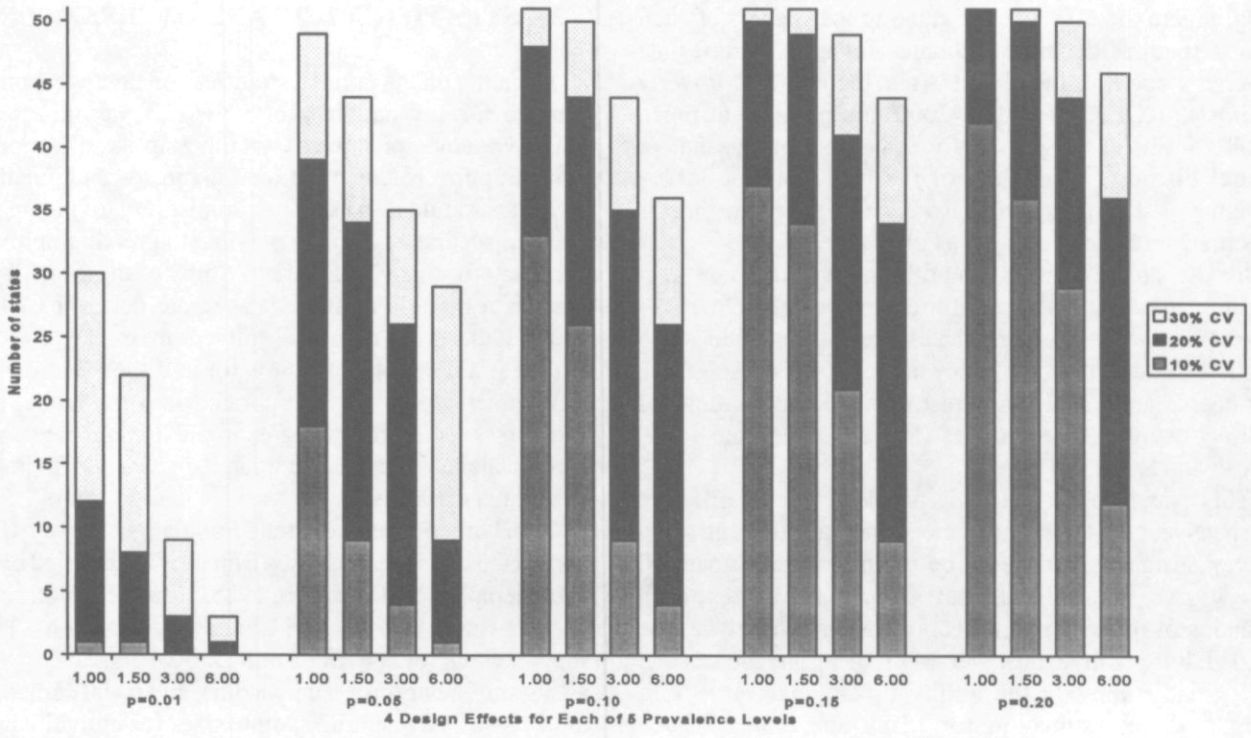

Figure 1. Number of States Meeting CV Criteria for 1995 NHIS (44,000 Households, 100,000 Individuals)

Stratification by small area assures a fixed sample size will be assigned to each small area, and thereby fixes the accuracy associated with direct estimates. Without such stratification, it may even be impossible to produce unbiased estimates for small areas that do contain some sample, because the probabilities of selection for sampled cases are a function of their entire stratum, both inside and outside the small area. For example, this can occur when part of a small area is in a stratum that crosses small area boundaries, and the sampled PSUs are in other small areas. To produce direct estimates requires either collapsing strata boundaries or small area boundaries.

By oversampling small areas it is possible to significantly improve the accuracy of direct estimates for these areas, while only incurring a minimal loss in accuracy for national estimates. As a simple example, consider a national survey with 5,000 respondents but where under a random sampling scheme 10 of the small areas would only receive 100 cases each. Alternatively one could double the sample size to 200 in each of these small areas while retaining the national sample size of 5,000. The effective sample size for national estimates would be reduced by this oversampling, but would remain more than 4,000, so the CV of national estimates would increase less than 10 percent. The CV for estimates in each of the 10 small areas would decrease 30 percent because the sample size was doubled.

Beginning in 1999 the U.S. National Household Survey on Drug Abuse has combined stratification and oversampling to produce direct estimates for every state (Chromy, Bowman and Penne 1999).

Singh *et al.* (1994) provided an example of oversampling small areas in the Canadian Labour Force Survey. Seventy percent of the sample was allocated to provide optimal national and provincial estimates. The remaining 30 percent were used to supplement small areas to improve their estimates. National CVs were increased between 10 and 20 percent by this compromise design, but unemployment insurance regions' estimates had CV reductions as large as 50 percent.

A similar design was used for the 2000 Danish Health and Morbidity Survey. The survey included two national samples, each of 6,000 respondents. An additional 8,000 respondents were distributed to assure that at least 1,000 respondents would be in each county.

The effect of oversampling on CVs can also be seen by comparing the 1996 CPS and 1995 NHIS with America's 1996 Survey of Income and Program Participation (SIPP). The CPS not only stratified by state, it also oversampled smaller states. The NHIS stratified by state but didn't oversample based on geography (minority groups were oversampled, but they tend to be located in the more populous states). In contrast, the SIPP did not stratify by state nor did it oversample. The ratio of the largest to smallest state sample size was 11:1 for CPS, 60:1 for SIPP, and 110:1 for NHIS. The corresponding ratio of CVs was 3.5:1, 7.5:1, and 10.5:1. Oversampling resulted in the CVs for the smallest states being reduced by almost a factor of two-thirds!

It is important to remember that oversampling based on geography doesn't necessarily reduce the variability in other domains of interest, for example demographic subgroups. The ratios of largest to smallest state sample sizes in the CPS were 15:1 for children, 20:1 for the elderly, 500:1 for Blacks, and 800:1 for Hispanics.

The 1994 U.S. National Employer Health Insurance Survey (NEHIS) oversampled smaller states to balance the need for accurate state and national estimates. The overall sample of 40,000 establishments had to be spread across all 51 states to provide direct estimates for all states. Three options were considered:

Option A: Optimal national allocation (based on total employment in the state) yielded very small sizes in some states.

Option B: Equal allocation to all states yielded inefficient national estimates.

Option C: Minimum 400 completes per state (allocate based on number of employees to the 0.3 power).

The corresponding ratio of largest to smallest state CVs were 7.2:1 for Option A, 1:1 for Option B, and 1.8:1 for Option C. Compared to Option C, the national CV with Option A was 17 percent lower, but with Option B was 22 percent higher. Option C was selected over Option A since it reduced the variation in state CVs by a factor of 4 while only moderately increasing the national CV.

3. COMBINING MULTIPLE YEARS

An inexpensive way to increase the sample sizes in small areas is to combine cycles of a repeated survey. Combining k years of an annual survey increases the effective sample size not quite k times. The reason for this is that usually consecutive years of the same survey are conducted in the same primary sampling units (PSUs) and even adjacent area segments. This results in some correlation between years, somewhat reducing the effective sample size.

One drawback to combining multiple years is that such estimates are slow to detect changes across time. If time series are a prime interest, alternative methods must be used to increase the sample size.

Table 2 shows for the 1995 NHIS how many states can achieve different levels of accuracy by aggregating across two or three years. Aggregation clearly helps achieve CVs of 30 and 20 percent. Even aggregating 3 years can't help many states achieve a CV of 10 percent.

Table 2
Summary of the Number of States (out of 51) That Have the Required 1995 NHIS Sample Size to Achieve a CV of 30-, 20-, and 10-Percent; Aggregating Multiple Years for Four Selected Variables (44,000 Households, 100,000 Individuals).

	Percent uninsured: all ages	Percent uninsured: under 19	Percent uninsured: low income children	Percent smokers: 18 and over
30-percent CV				
1 year	42	31	28	45
2 years	46	35	36	50
3 years	49	41	37	51
20-percent CV				
1 year	31	13	10	36
2 years	36	29	24	44
3 years	42	31	31	46
10-percent CV				
1 year	7	2	2	14
2 years	14	3	3	25
3 years	22	7	4	32

4. HARMONIZATION ACROSS SURVEYS

Harmonizing questions across surveys is another inexpensive way to improve estimation. Eurostat has been making a major effort to harmonize a number of surveys both between countries and within. The European Community Household Panel Survey (ECHP) is an attempt to collect consistent information across the member countries. Similar standardization is ongoing in each country's Labour Force Survey. This harmonization across countries improves international comparisons.

Harmonizing across surveys of the same population increases sample sizes, improving small area estimates. Statistics Finland has been harmonizing the process for collecting income and other variables in its surveys. The Permanent Survey on Living Conditions (POLS) at Statistics Netherlands uses a common procedure for collecting basic information in a series of social surveys.

Even if the questionnaire wording is consistent across surveys, the data may not be completely comparable. Different modes of data collection can cause differences, as can the placement of questions (Groves 1989).

5. DUAL-FRAME ESTIMATION

In some situations it is possible to supplement an in-person survey with telephone data collection, thereby increasing the sample size in a small area at more limited expense. The Dutch Housing Demand Survey is a national in-person survey. To produce small area estimates telephone supplementation is used in over 100 municipalities. Table 3 shows the size of the national in-person survey, telephone supplement, and total sample in ten selected municipalities.

Table 3
Dual-Frame Completes for Municipalities in the Dutch Housing Demand Survey

Municipality	In-Person National Survey	Telephone Supplement	Total
Leek	56	569	625
Marum	29	299	328
Slochteren	44	456	500
Zuidhorn	54	558	612
Emmen	770	224	994
Avereest	134	465	599
Bathmen	24	506	530
Dalfsen	157	466	623
Deventer	316	335	651
Diepenveen	47	336	383

Sirken and Marker (1993) described dual-frame estimation for the U.S. National Health Insurance Survey (NHIS) based on its 1985-94 design. Table 4 examines the same idea for the current design implemented beginning in 1995. The table compares the ability to produce state estimates with national in-person survey interviews and with unbiased dual-frame estimation using an unlimited number of supplemental telephone interviews. (Up to 100, 200, and 2,000 telephone interviews per state are required to achieve CVs of 30-, 20-, and 10- percent, respectively.) When a small area has a large percentage of households without telephones, no amount of telephone supplementation may be sufficient to achieve unbiased estimates with the desired accuracy.

In such situations, it may only be possible to achieve a desired level of accuracy using a potentially biased estimator that combines all data regardless of the mode of collection. The relative root mean square error (RRMSE) must then be used instead of the CV to measure accuracy. However, for some characteristics households with telephones have different expectations than households without telephones. In such situations the bias can again prevent achieving the desired accuracy. The bias for each of these variables was estimated by comparing NHIS responses from households with and without telephones. Table 5 shows how the number of states for which a 10 percent RRMSE can be achieved varies by question, a function of the bias in telephone households and the telephone penetration rate in each state.

Small areas with high telephone penetration rates, for characteristics with different expectations for telephone and non-telephone households, are better able to produce accurate estimates using an unbiased dual-frame estimator. Small areas with lower penetration rates, for characteristics with similar telephone and non-telephone households, produce more accurate estimates with a potentially biased dual-frame estimator. Using the appropriate dual-frame estimator for a given small area and characteristic can allow accurate estimates to be produced for a large percentage of small areas.

Table 4
The Number of States Able to Achieve 30-, 20-, 10-Percent CV With the 1995 NHIS Area Sample Only, With Unbiased Dual-Frame Estimation Using a RDD Supplement, or not at All, for Four Specific Variables

CV	Data sources	Percent uninsured: all ages	Percent uninsured: under 19	Percent uninsured: low income children	Percent smokers: 18 and over
30%	With area sample only	42	31	31	46
	With RDD supplement	9	20	19	5
	Unable to meet requirement	0	0	1	0
20%	With area sample only	32	15	10	37
	With RDD supplement	19	35	40	14
	Unable to meet requirement	0	1	1	0
10%	With area sample only	8	2	2	15
	With RDD supplement	40	41	39	36
	Unable to meet requirement	3	8	10	0

Table 5
The Number of States Able to Achieve 10-Percent RRMSE With the 1995 NHIS Area Sample Only, With a RDD Supplement, or not at all, for the Four Specific Variables

Data source	Percent uninsured: all ages	Percent uninsured: under 19	Percent uninsured: low income children	Percent smokers: 18 and over
With area sample only	8	2	2	15
With RDD supplement				
Unbiased Estimator	40	41	39	36
Biased Estimator	30	47	49	35
Unable to meet requirement				
Unbiased Estimator	3	8	10	0
Biased Estimator	13	2	0	1

6. IMPROVING POINT AND VARIANCE ESTIMATION

When sufficient sample size exists to produce small area estimates there are additional steps that can be taken to improve their accuracy. SIPP does not stratify by state, to improve state estimates it reweights the estimates to control totals at the state level. This is very important when the stratification doesn't match the analytic domains. The use of control totals also improves subpopulation (*e.g.*, demographic) size estimates for the small areas. However, it is not possible to control as many subpopulations in a small area as can be done at the national level, due to the smaller sample sizes.

There are also many techniques to improve variance estimation for small areas. Typically there will be very few sampled PSUs in a given small area. This provides few degrees of freedom for estimating between-PSU (or total) variance. One solution is to average estimates of variance across small areas, but this covers up the fact that estimates are generally much better for some areas than for others. Alternatively generalized variance functions (GVFs) can be used to smooth variance estimates.

A preferable solution is to address small area variance estimation at the design stage. Increasing the number of PSUs, with a corresponding reduction in sample size in each PSU, can significantly improve both point and variance estimation, often at little extra cost. Singh *et al.* (1994) suggested increasing the number of PSUs to control sample sizes in unplanned small areas. Remembering Fuller's observation that "The client will always require more than is specified at the design stage," it is impossible to anticipate all small areas of interest. By having more PSUs the likelihood is increased that actual data will have been collected from unplanned analytic domains.

Kalton (1994) suggested a second reason for increasing the number of PSUs. His concern was that more PSUs per small area would greatly increase the stability of variance estimates. This is true even in very large national surveys with many PSUs. The NHIS was redesigned in 1995 increasing the number of PSUs from 196 to 359. Of these 359 PSUs 264 were noncertainty PSUs. This still resulted in only 7 states having more than 8 noncertainty PSUs. While direct variance estimation for individual states is still problematic for most states, there is an increased opportunity to develop average variance estimates for groups of states with common characteristics, rather than having to group all states together in a national average.

7. SUMMARY

There will always be a need for indirect small area estimation methods since the entire set of analytic domains is never known in advance. This need for small area estimates is growing around the world. There are, however, many actions that can be taken at the design stage to improve direct small area estimates, both point estimates and variance estimates. These steps include stratification consistent with known analytic domains, oversampling smaller areas, and increasing the number of PSUs. Given the data it is often possible to combine data from multiple years, from other surveys with whom questions have been harmonized, and through dual-frame estimation techniques. These steps will both reduce the need for indirect estimates and improve the accuracy of those estimates when they are required.

REFERENCES

CHROMY, J.R., BOWMAN, K.R. and PENNE, M.A. (1999). The National Household Survey on Drug Abuse Sample Design Plan. Prepared for the Substance Abuse and Mental Health Services Administration, Rockville Maryland.

CITRO, C.F., and KALTON, G. (2000). Small-area Estimates of School-age Children in Poverty: Evaluation of Current Methodology. National academy press, Washington, D.C.

FULLER, W.A. (1999). Environmental surveys over time. *Journal of agricultural, Biological, and Environmental Statistics*, 4, 331-345.

GHOSH, M., and RAO, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55-93.

GROVES, R.M. (1989). *Survey Errors and Survey Costs*. New York: John Wiley & Sons.

KALTON, G. (1994). Comments on Singh, Gambino and Mantel. *Survey Methodology*, 20, 18-20.

MARKER, D.A. (1999). Organization of small area estimators using a generalized linear regression framework. *Journal of Official Statistics*, 15, 1-24.

RAO, J.N.K. (1999). Some recent advances in model-based small area estimation. *Survey Methodology*, 25, 175-186.

SINGH, M.P., GAMBINO, J. and MANTEL, H.J. (1994). Issues and strategies for small area data. *Survey Methodology*, 20, 3-14.

SIRKEN, M.G., and MARKER, D.A. (1993). Dual frame sample surveys based on NHIS and state RDD surveys. *Proceedings of the 1993 Public Health Conference on Records and Statistics.*

Survey Methodology, December 2001
Vol. 27, No. 2, pp. 189-196
Statistics Canada

A Repeated Half-Sample Bootstrap and Balanced Repeated Replications for Randomly Imputed Data

HIROSHI SAIGO, JUN SHAO and RANDY R. SITTER[1]

ABSTRACT

In this paper, we discuss the application of the bootstrap with a re-imputation step to capture the imputation variance (Shao and Sitter 1996) in stratified multistage sampling. We propose a modified bootstrap that does not require rescaling so that Shao and Sitter's procedure can be applied to the case where random imputation is applied and the first-stage stratum sample sizes are very small. This provides a unified method that works irrespective of the imputation method (random or nonrandom), the stratum size (small or large), the type of estimator (smooth or nonsmooth), or the type of problem (variance estimation or sampling distribution estimation). In addition, we discuss the proper Monte Carlo approximation to the bootstrap variance, when using re-imputation together with resampling methods. In this setting, more care is needed than is typical. Similar results are obtained for the method of balanced repeated replications, which is often used in surveys and can be viewed as an analytic approximation to the bootstrap. Finally, some simulation results are presented to study finite sample properties and various variance estimators for imputed data.

KEY WORDS: Hotdeck; Percentile method; Monte Carlo; Imputation; Bootstrap sample size.

1. INTRODUCTION

Item nonresponse is a common occurrence in surveys and is usually handled by imputing missing item values. The various imputation methods used in practice can be classified into two types: deterministic imputation, such as mean, ratio and regression imputation, typically using the respondents and some auxiliary data observed on all sampled elements; and random imputation. In both cases the imputation is often applied within imputation classes formed on the basis of auxiliary variables. This article focuses on random imputation.

Typically, random imputation is done in such a way that applying the usual estimation formulas to the imputed data set produces asymptotically unbiased and consistent survey estimators (*e.g.*, means, totals, quantiles). More details about random imputation are provided in section 2. It is common practice to also treat the imputed values as true values when estimating variances of survey estimators. This leads to serious underestimation of variances if the proportion of missing data is appreciable, and to poor confidence intervals.

There have been some proposals in the literature to circumvent this difficulty. For random imputation, Rubin (1978) and Rubin and Schenker (1986) proposed the multiple imputation method to account for the inflation in the variance, which can be justified from a Bayesian perspective (Rubin 1987). Adjusted jackknife methods for variance estimation have been proposed for both random and deterministic imputations (Rao and Shao 1992; Rao 1993; Rao and Sitter 1995; Sitter 1997), under stratified multistage sampling. However, it is well known that the jackknife cannot be applied to non-smooth estimators, *e.g.*, a sample quantile or an estimated low income proportion (Mantel and Singh 1991).

There are two methods available for handling randomly imputed data for both smooth and non-smooth estimators: the adjusted balanced repeated replication (BRR) methods proposed by Shao, Chen and Chen (1998); and the bootstrap method proposed by Shao and Sitter (1996) (see also Efron 1994) with a re-imputation step to capture the imputation variance. The bootstrap method is more computer intensive but is easy to motivate and understand, and provides a unified method that works irrespective of the imputation method (random or nonrandom), the type of $\hat{\theta}$ (smooth or nonsmooth), or the type of problem (variance estimation or sampling distribution estimation).

In this article we continue the work by Shao and Sitter (1996). First, we show in section 3 how Shao and Sitter's bootstrap procedure can be modified to handle very small stratum sizes (*e.g.*, two psu's per stratum). Second, we discuss in section 4 the proper Monte Carlo approximation to the bootstrap estimators, a problem for which more care is needed when random re-imputation is applied than is typical. This has no detrimental effect on bootstrap confidence intervals based on the percentile method, but if done incorrectly, will cause the bootstrap-t to perform poorly. Third, we consider a BRR variance estimation method with a re-imputation step, which can be viewed as an analytic and symmetric approximation to the bootstrap method. Finally, we present some simulation results to study properties of various bootstrap and BRR variance estimators.

[1] Hiroshi Saigo, School of Political Science and Economics, Waseda University, 1-6-1 Nishiwaseda Shinjuku, Tokyo, 169-8050 Japan; Jun Shao, Professor, Department of Statistics, University of Wisconsin, Madison, WI 53706, USA; Randy R. Sitter, Professor, Department of Statistics and Acturial Science, Simon Fraser University, Burnaby, BC, Canada V5A 1S6.

2. STRATIFIED MULTISTAGE SAMPLING AND RANDOM IMPUTATION

Though the methods discussed in this article can be more generally applied, we restrict attention to the commonly used stratified multistage sampling design. Suppose that the population contains H strata and in stratum h, n_h clusters are selected with probabilities p_{hi}, $i = 1, ..., n_h$. Samples are taken independently across strata. In the case of complete response on item y, let

$$\hat{Y}_h = \sum_{i=1}^{n_h} \hat{Y}_{hi} / (n_h p_{hi})$$

be a linear unbiased estimator of the stratum total Y_h, where $\hat{Y}_{hi}$ is a linear unbiased estimator of the cluster total Y_{hi} for a selected cluster based on sampling at the second and subsequent stages. A linear unbiased estimator of the total, $Y = \Sigma Y_h$, is given by $\hat{Y} = \Sigma \hat{Y}_h$, which may be written as

$$\hat{Y} = \sum_{(hik) \varepsilon s} w_{hik} y_{hik}, \qquad (1)$$

where s is the complete sample of elements, and w_{hik} and y_{hik} respectively denote the sampling weight and the item value attached to the (hik)-th sampled element.

Often a survey estimator, $\hat{\theta}$, can be expressed as a function of a vector of estimated totals as in (1). If one is interested in the population distribution function, it can be estimated by $\hat{F}_n(t) = \Sigma_s w_{hik} I(y_{hik} \le t) / \hat{U}$, where $I(\cdot)$ is the usual indicator function and $\hat{U} = \Sigma_s w_{hik}$. Some non-smooth estimators that are of interest are the p-th sample quantile, $\hat{F}^{-1}(p)$, where $\hat{F}^{-1}$ is the quantile function of $\hat{F}$, and the sample low income proportion $\hat{F}[(1/2)\hat{F}^{-1}(1/2)]$.

Suppose that the value y_{hik} is observed for $(hik) \varepsilon s_r \subset s$, termed a respondent, while for others, $(hik) \varepsilon s_m$, it is missing, termed a nonrespondent, with $s = s_r \cup s_m$. When there are missing data, it is common practice to use $\{y_{hik}: (hik) \varepsilon s_r\}$ to obtain imputed values $\bar{y}_{hik}$ for $(hik) \varepsilon s_m$ and then treat these imputed values as if they were true observations and estimate Y with

$$\hat{Y}_I = \sum_{s_r} w_{hik} y_{hik} + \sum_{s_m} w_{hik} \bar{y}_{hik}. \qquad (2)$$

In practice, the accuracy of the imputation is improved by first forming several imputation classes using control variables observed on the entire sample, and then imputing within imputation class. For simplicity we consider a single imputation class.

Random imputation entails imputing the missing data by a random sample from the respondents, or, in the presence of auxiliary data, by using a random sample of residuals. If the imputation is suitably done, the estimator $\hat{Y}_I$ in (2) is asymptotically unbiased and consistent, although it is not as efficient as $\hat{Y}$ in (1). Throughout this article, we assume that, either

> within each imputation cell, the response probability for a given variable is a constant, the response statuses for different units are independent, and imputation is carried out within each imputation cell and independently across the imputation cells,

or

> within each imputation cell, the response probability of a given variable does not depend on the variable itself (but may depend on the covariates used for imputation), imputation is carried out independently across the imputation cells, and within an imputation cell, imputation is performed according to a model that relates the variable being imputed to the covariates used for imputation.

We also assume the same asymptotic setting as that in Shao *et al.* (1998). Thus, consistency (or asymptotic unbiasedness) refers to convergence of estimators (or expectations of estimators) under the assumption in Shao, *et al.* (1998), as the first-stage sample size $n = \Sigma n_h$ increases to infinity.

There are many methods of random imputation. We consider only two in this article: the weighted hotdeck considered in Rao and Shao (1992), which we refer to simply as random imputation, and the adjusted weighted hotdeck proposed in Chen, Rao and Sitter (2000), which we refer to as adjusted random imputation. Our results can be easily extended to random imputation with residuals in the presence of auxiliary data (*e.g.*, random regression imputation). Generalizations to other types of random imputation may be possible, but will not be considered here.

Random imputation randomly selects donors, $\bar{y}_{hik}$ from $\{y_{hik}: (hik) \varepsilon s_r\}$ with replacement with probabilities $w_{hik}/\hat{T}$, where $\hat{T} = \Sigma_{s_r} w_{hik}$. In this case $E_I(\hat{Y}_I) = (\hat{S}/\hat{T})\hat{U} = \hat{Y}_r$, a ratio estimator which is asymptotically unbiased and consistent for Y, where $\hat{S} = \Sigma_{s_r} w_{hik} y_{hik}$. Here E_I denotes expectation under the random imputation. The variance of $\hat{Y}_I$ is larger than the variance of $\hat{Y}_r$ because of the random imputation. However, the distribution of item values in the imputed data set is preserved.

Adjusted random imputation simply uses $\tilde{\eta}_{hik} = \bar{y}_{hik} + (\hat{S}/\hat{T} - \tilde{S}/\tilde{T})$ as the imputed values instead of $\bar{y}_{hik}$, where $\tilde{S} = \Sigma_{s_m} w_{hik} \bar{y}_{hik}$, $\tilde{T} = \Sigma_{s_m} w_{hik}$ and $\bar{y}_{hik}$ are the imputed values from random imputation. Chen *et al.* (2000) show that this method completely eliminates the variability due to the random imputation for estimating the population total. That is $\tilde{Y}_I = \Sigma_{s_r} w_{hik} y_{hik} + \Sigma_{s_m} w_{hik} \eta_{hik} = \hat{Y}_r$. The method also retains the distribution of item values in the imputed data set. However, the resulting imputed values need not be actual realizations.

An imputed estimator of the distribution function under random imputation is given by

$$\hat{F}_I(t) = \left[\sum_{s_r} w_{hik} I(y_{hik} \le t) + \sum_{s_m} w_{hik} I(\bar{y}_{hik} \le t) \right] / \hat{U}. \qquad (3)$$

An imputed estimator of the distribution function under adjusted random imputation, denoted $\tilde{F}_I(t)$, is simply obtained by replacing $\bar{y}_{hik}$ in (3) by $\tilde{\eta}_{hik}$. For estimating the

distribution function, adjusted random imputation does not eliminate the imputation variance as it does for estimating the total. However, Chen *et al.* (2000) show that it does significantly reduce the imputation variance when compared to random imputation. Both $\hat{F}_I(t)$ and $\tilde{F}_I(t)$ are asymptotically unbiased and consistent.

For studying variance estimation with resampling methods, we assume that n/N is negligible, where $n = \Sigma n_h, N = \Sigma N_h$ and N_h is the number of first-stage clusters in the population.

3. A REPEATED HALF-SAMPLE BOOTSTRAP

When there are imputed missing data, naive bootstrap variance estimators obtained by treating the imputed data set, $\boldsymbol{Y}_I$, as $\boldsymbol{Y} = \{y_{hik}: (hik) \in s\}$, the data set of no missing values, do not capture the inflation in variance due to imputation and/or missing data and lead to serious underestimation. As a result, they are inconsistent. This is so, because simply treating $\boldsymbol{Y}_I$ as $\boldsymbol{Y}$ ignores the imputation process. This was noted by Shao and Sitter (1996) and they proposed re-imputing the bootstrap data set in the same way as the original data set was imputed. The bootstrap procedure in Shao and Sitter (1996) can be described as follows.

1. Draw a simple random sample $\{y^*_{hi}: i = 1, ..., n_h - 1\}$ with replacement from the sample $\{\tilde{y}_{hi}: i = 1, ..., n_h\}$, $h = 1, ..., H$, independently across the strata, where $\tilde{y}_{hi} = \{y_{hij}: (h, i, j) \in s_r\} \cup \{\tilde{y}_{hij}: (h, i, j) \in s_m\}$.
2. Let a^*_{hij} be the response indicator associated with y^*_{hij}, $s^*_m = \{(h, i, j): a^*_{hij} = 0\}$ and $s^*_r = \{(h, i, j): a^*_{hij} = 1\}$. Apply the same imputation procedure used in constructing the imputed data set $\boldsymbol{Y}_I$ to the "nonrespondents" in s^*_m, using the "respondents" in s^*_r. Denote the bootstrap analogue of $\boldsymbol{Y}_I$ by $\boldsymbol{Y}^*_I$.
3. Obtain the bootstrap analogue $\hat{\theta}^*_I$ of $\hat{\theta}$, based on the imputed bootstrap data set $\boldsymbol{Y}^*_I$. For example, if $\hat{\theta} = \hat{Y}$ in (1) and $\hat{\theta}_I = \hat{Y}_I$ in (2), then

$$\hat{\theta}^*_I = \hat{Y}^*_I = \sum_{s^*_r} w^*_{hik} y^*_{hik} + \sum_{s^*_m} w^*_{hik} \tilde{y}^*_{hik}, \tag{4}$$

where $\tilde{y}^*_{hik}$ is the imputed value using the bootstrap data and w^*_{hik} is $n_h/(n_h - 1)$ times the survey weight associated with y^*_{hik} (to reflect the fact that the bootstrap sample size is $n_h - 1$, not n_h). The bootstrap estimator of $\mathrm{Var}(\hat{\theta}_I)$ is

$$v_B(\hat{\theta}_I) = \mathrm{Var}^*(\hat{\theta}^*_I), \tag{5}$$

where Var^* is the conditional variance with respect to $\boldsymbol{Y}^*_I$, given $\boldsymbol{Y}_I$.

Shao and Sitter (1996) show that the bootstrap estimator defined in (5) is consistent for both smooth and nonsmooth estimators $\hat{\theta}$. When a random imputation method is considered, an implicit condition in their development is that $n_h/(n_h - 1)$ goes to 1. This can be seen from the special case of $\hat{\theta} = \hat{Y}$. From (2),

$$\mathrm{Var}(\hat{Y}_I) = \mathrm{Var}\left[E_I(\hat{Y}_I)\right] + E\left[\mathrm{Var}_I(\hat{Y}_I)\right]$$

$$= \mathrm{Var}\left(\frac{\sum_{s_r} w_{hik} y_{hik} \sum_{s} w_{hik}}{\sum_{s_r} w_{hik}}\right) + E\left(\hat{\sigma}^2 \sum_{s_m} w^2_{hik}\right), \tag{6}$$

where

$$\hat{\sigma}^2 = \sum_{s_r} w_{hik}(y_{hik} - \bar{y}_r)^2 \Big/ \sum_{s_r} w_{hik},$$

$$\bar{y}_r = \sum_{s_r} w_{hik} y_{hik} \Big/ \sum_{s_r} w_{hik}.$$

Similarly, by (4),

$$\mathrm{Var}^*(\hat{Y}^*_I) = \mathrm{Var}^*\left(\frac{\sum_{s^*_r} w^*_{hik} y^*_{hik} \sum_{s^*} w^*_{hik}}{\sum_{s^*_r} w^*_{hik}}\right)$$

$$+ E^*\left(\hat{\sigma}^{*2} \sum_{s^*_m} w^{*2}_{hik}\right), \tag{7}$$

where

$$\hat{\sigma}^{*2} = \sum_{s^*_r} w^*_{hik}(y^*_{hik} - \bar{y}^*_r)^2 \Big/ \sum_{s^*_r} w^*_{hik},$$

$$\bar{y}^*_r = \sum_{s^*_r} w^*_{hik} y^*_{hik} \Big/ \sum_{s^*_r} w^*_{hik}.$$

From the theory of the bootstrap, the first terms on the right hand side of (6) and (7) converge to the same quantity, as do $\hat{\sigma}^2$ and $\hat{\sigma}^{*2}$. Thus, Shao and Sitter's bootstrap is consistent if $\Sigma_{s^*_m} w^{*2}_{hik}$ and $\Sigma_{s_m} w^2_{hik}$ converge to the same quantity, which is true if $n_h/(n_h - 1)$ converges to 1 for all h, because

$$E^*\left(\sum_{s^*_m} w^{*2}_{hik}\right) = E^*\left[\sum_{s^*} (1 - a^*_{hik}) w^{*2}_{hik}\right]$$

$$= \sum_{s} (1 - a_{hik}) w^2_{hik} n_h/(n_h - 1).$$

The second term on the right hand side of (6) is the variance component corresponding to random imputation, which is typically a small portion of the overall variance. Thus, the overestimation due to $n_h/(n_h - 1)$ is serious only when the n_h's are very small. The case $n_h = 2$ is, however, an important special case.

We now propose a bootstrap method which has no difficulty in the case of very small n_h's while remaining valid more generally. Note that the use of bootstrap sample size $n_h - 1$ is to ensure that the first term on the right hand side of (7) has the same limit as the first term on the right

hand side of (6) (Rao and Wu 1988). When n_h is used as the bootstrap sample size in stratum h, Rao and Wu (1988) showed that in the case of no missing data, the bootstrap variance estimator underestimates. They proposed a rescaling to circumvent the problem, but rescaling does not produce correct bootstrap estimators in the presence of imputed data.

What is ideally required for our problem is a bootstrap method with the bootstrap sample size equal to the original sample size n_h which produces an asymptotically unbiased variance estimator (in the case of no missing data) without rescaling. We now show that this can be accomplished as follows. Suppose that there is no missing data and that all of the $n_h = 2m_h$'s are even. Take a simple random sample of size m_h without replacement independently from $\{ y_{hi}: i = 1, ..., n_h \}$ and repeat each obtained unit twice to get $\{ y^*_{hi}: i = 1, ..., n_h \}$. We call this method the repeated half-sample bootstrap. The resulting v_B will then be approximately unbiased and consistent. In the linear case where $\hat{Y} = \sum_{(hik)} w_{hik} y_{hik} = \sum_h \sum_{i=1}^{n_h} y_{hi}/n_h = \sum_h \bar{y}_h$ and $y_{hi} = \sum_{k=1}^{n_{hi}} n_h w_{hik} y_{hik}$, the consistency of v_B follows from

$$\begin{aligned}
\mathrm{Var}^*(\hat{Y}^*) &= \sum_h \mathrm{Var}^*(\bar{y}^*_h) = \sum_h \mathrm{Var}^*\left(\frac{1}{n_h} \sum_{i=1}^{n_h} y^*_{hi} \right) \\
&= \sum_h \mathrm{Var}^*\left(\frac{2m_h}{n_h} \frac{1}{m_h} \sum_{i=1}^{m_h} y^*_{hi} \right) \\
&= \sum_h \mathrm{Var}^*\left(\frac{1}{m_h} \sum_{i=1}^{m_h} y^*_{hi} \right) \\
&= \sum_h \frac{(1 - \tfrac{1}{2})}{m_h} \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2 \\
&= \sum_h s_h^2 / n_h,
\end{aligned}$$

the usual approximately unbiased and consistent estimator of variance, where $s_h^2 = (n_h - 1)^{-1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2$. The consistency of v_B for a nonlinear $\hat{\theta}_I$ follows from the linear case and Taylor's expansion, when $\hat{\theta}_I$ is a function of weighted averages, or the arguments used in Shao and Rao (1994), Shao and Sitter (1996), and Shao *et al.* (1998) when $\hat{\theta}_I$ is non-smooth such as a median.

If $n_h = 2m_h + 1$ is odd, it is not possible to take an exact half-sample. In this case, the following two results lead us to an adaptation of the above idea:

i) If we choose a simple random resample of size $m_h = (n_h - 1)/2$ without replacement and repeat each unit twice, we end up with $n_h - 1$ units. If we obtain an additional unit by selecting one at random from the $n_h - 1$ units already resampled, $\mathrm{Var}^*(\hat{Y}^*) = \sum_h (n_h + 3) s_h^2 / n_h^2$;

ii) If we choose a simple random resample of size $m_h + 1$ without replacement and repeat each unit twice, we end up with $n_h + 1$ units. If we discard one of these at random, $\mathrm{Var}^*(\hat{Y}^*) = \sum_h (n_h - 1) s_h^2 / n_h^2$.

Thus, if we used method (i) with probability 1/4 and method (ii) with probability 3/4 at each bootstrap replication, we obtain the desired result. This repeated half-sample bootstrap method yields approximately unbiased variance estimates without rescaling and has a bootstrap sample size equal to the original sample size. Thus, if we use this bootstrap for Step 1 of the method of Shao and Sitter (1996) as described above, the resulting bootstrap estimators are asymptotically unbiased and consistent for any n_h, under the regularity conditions stated in Shao and Sitter (1996) and Shao *et al.* (1998).

4. THE PROPER MONTE CARLO FOR THE BOOTSTRAP

If v_B in (5) has no explicit form, one may use the Monte Carlo approximation

$$v_B(\hat{\theta}_I) \approx \frac{1}{B} \sum_{b=1}^{B} (\hat{\theta}^*_{I(b)} - \bar{\theta}^*_I)^2, \tag{8}$$

where $\bar{\theta}^*_I = B^{-1} \sum_{b=1}^{B} \hat{\theta}^*_{I(b)}$, $\hat{\theta}^*_{I(b)} = \hat{\theta}(Y^*_{I(b)})$, and $Y^*_{I(b)}$, $b = 1, ..., B$, are independent re-imputed bootstrap data sets. It is common practice in many applications of the bootstrap to replace the average of the bootstrap estimators $\bar{\theta}^*_I$ in (8) by the original estimator $\hat{\theta}_I$ (see Rao and Wu 1985, page 232). The latter is simpler to use and is thus the most common. With no imputed data, this is usually correct. However, using the analogue with the re-imputed bootstrap is not correct. The reason is that $\hat{\theta}_I$ is the result of a single realization of the random imputation, while $\bar{\theta}^*_I \approx E^*(\hat{\theta}^*_I) \approx E_I(\hat{\theta}_I)$ since we are averaging over repeated re-imputations, and $\hat{\theta}_I$ and $E_I(\hat{\theta}_I)$ are not close for random imputation. When $\hat{\theta}_I = \hat{Y}_I$, for example, $E_I(\hat{Y}_I) = \hat{Y}_r$ given in section 2 and the difference $\hat{Y}_I - \hat{Y}_r$ is not a relatively negligible term when random imputation is used. Thus,

$$v_{B2} = \frac{1}{B} \sum_{b=1}^{B} (\hat{\theta}^*_{I(b)} - \hat{\theta}_I)^2 = \frac{1}{B} \sum_{b=1}^{B} (\hat{\theta}^*_{I(b)} - \bar{\theta}^*_I)^2 + (\bar{\theta}^*_I - \hat{\theta}_I)^2$$

and the first term goes to $\mathrm{Var}^*(\hat{\theta}^*_I)$ as $B \to \infty$ but the second term does not go to zero which implies that v_{B2} badly overestimates the variance. This is not only true for the proposed repeated half-sample bootstrap but also for those considered in Shao and Sitter (1996).

One should also note that using the $\hat{\theta}^*_{I(b)}$, $b = 1, ..., B$ to obtain bootstrap confidence intervals via the percentile method avoids this concern since the histogram of these values will be correctly centered about $E^*(\hat{\theta}^*_I)$. However, one must take more care with bootstrap-t confidence

intervals. It is important that one define $t_b^* = (\hat{\theta}_{I(b)}^* - \bar{\theta}_{I(.)}^*)/\sigma_b^*$ (not $t_b^* = (\hat{\theta}_I^{*b} - \hat{\theta}_I)/\sigma_b^*$) and use $\{\hat{\theta}_I - t_U^*\sigma_b^*, \hat{\theta}_I - t_L^*\sigma_b^*\}$, where $\sigma_b^{*2} = v_B(Y_I^{*b})$, $t_L^* = \hat{\text{CDF}}_t^{-1}(\alpha)$, $t_U^* \hat{\text{CDF}}_t^{-1}(1-\alpha)$ and $\hat{\text{CDF}}_t(x) = \#\{t_b^* \le x; b = 1, ..., B\}/B$.

5. A REPEATED BRR

We first describe the most common application of the BRR, $n_h = 2$ clusters per stratum (McCarthy 1969) in the setting of no missing data. A set of B balanced half-samples or replicates is formed by deleting one first-stage cluster from the sample in each stratum, where this set is defined by a $B \times H$ matrix $(\delta_{bh})_{B\times H}$ with $\delta_{bh} = +1$ or -1 according to whether the first or the second first-stage cluster of stratum h is in the b-th half-sample and $\sum_{r=1}^{B} \delta_{bh}\delta_{bh'} = 0$ for all $h \neq h'$; that is, the columns of the matrix are orthogonal. A minimal set of B balanced half-samples can be constructed from a $B \times B$ Hadamard matrix by choosing any H columns excluding the column of all +1's, where $H + 1 \le B \le H + 4$. Let $\hat{\theta}_{(b)}$ be the survey estimator computed from the b-th half-sample. The estimator $\hat{\theta}_{(b)}$ can be obtained using the same formula as for $\hat{\theta}$ with w_{hik} changed to $w_{hik(b)}$, which equals $2w_{hik}$ or 0 according to whether or not the (hi)-th cluster is selected in the b-th half-sample or not. The BRR variance estimator for $\hat{\theta}$ is then given by

$$v_{\text{BRR}} = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}_{(b)} - \bar{\theta}_{(.)}\right)^2, \qquad (9)$$

where $\bar{\theta}_{(.)} = \Sigma_b \hat{\theta}_{(b)}/B$, and is often replaced by $\hat{\theta}$. The variance estimator v_{BRR} has been shown to be consistent for smooth functions of estimated totals by Krewski and Rao (1981) and for nonsmooth estimators by Shao, and Wu (1992) and Shao and Rao (1994).

A naive BRR for problems with randomly imputed data would be obtained as in (9) with $\hat{\theta}_{(b)}$ and $\bar{\theta}_{(.)}$ replaced by $\hat{\theta}_{I(b)}$ and $\bar{\theta}_{(.)} = B^{-1}\Sigma_b \hat{\theta}_{I(b)}$, where $\hat{\theta}_{I(b)}$ is the estimator calculated from Y_I using the BRR weights. But this produces inconsistent variance estimators because it fails to take into account the effect of missing data and the random imputation.

To correctly apply the BRR in the presence of random imputation by using re-imputation, we must deal with the issue of n_h being small. Recall that for the bootstrap such small n_h's caused difficulty because the stratum resample size, $n_h - 1$, was smaller than the original stratum sample size, n_h. This is true for the BRR, as well. We propose an easy way to circumvent this difficulty. Rather than obtaining the b-th BRR replicate of the estimator, $\hat{\theta}_{(b)}$, from the same formula as for $\hat{\theta}$ but with weights $w_{hik(b)}$ equal $2w_{hik}$ or 0 according as to whether the (hi)-th cluster is selected in the b-th half-sample or not, instead use the original weights but include the (hi)-th cluster twice or not at all according as to whether the (hi)-th cluster is selected in the b-th half-sample or not. If we view the BRR in this way: i) the resulting v_{BRR} in (9) remains the same; and ii) the resample size is the same as the original sample size. This repeated BRR can be viewed as a type of balanced bootstrap, however one should note that the balanced bootstrap described in Nigam and Rao (1996) for the case of no missing data does not work in this case because, though it uses a resample size $n_h = 2$ in each stratum, it does so in such a way as to still require rescaling and thus will not work in the presence of random imputation.

The proposed repeated BRR has no difficulty in the presense of random imputation. The procedure becomes

1. Form the set of half-samples, 1 unit per stratum, using a Hadamard matrix as described above.
2. Obtain the b-th BRR replicate by repeating each unit in the obtained half-sample twice. Denote this $\{y_{hi}^*: i = 1, ..., n_h = 2\}$.
3. Let a_{hij}^* be the response indicator associated with y_{hij}^*, $s_m^* = \{(h,i,j): a_{hij}^* = 0\}$, and $s_r^* = \{(h,i,j): a_{hij}^* = 1\}$. Apply the same imputation procedure used in constructing Y_I to the units in s_m^*, using the "respondents" in s_r^*. Denote the b-th BRR replicate of Y_I by $Y_{I(b)}^*$.
4. Obtain the BRR analogue $\hat{\theta}_{I(b)}^*$ of $\hat{\theta}$, based on the imputed BRR data set $Y_{I(b)}^*$.
4. Repeat 1-4 for each row of the $B \times H$ matrix to get $\hat{\theta}_{I(b)}^*$ for $b = 1, ..., B$ and apply the standard BRR formula (9) to obtain BRR variance estimators for $\hat{\theta}_I$, with $\bar{\theta}_{(.)} = B^{-1}\Sigma_b \hat{\theta}_{I(b)}^*$ (For the same reason that is discussed in section 4, we should not replace $\bar{\theta}_{I(.)}$ by $\hat{\theta}_I$).

We can extend this idea to cases with $n_h > 2$ by using the same strategy with half-samples obtained from balanced orthogonal multi-arrays (BOMA's) (Sitter 1993). For example, Table 1 gives a set of $B = 24$ balanced resamples for $H = 7$ strata with $n_h = 4$ psu's in each stratum. It is derived using the BOMA given in Table 1 of Sitter (1993) and repeating each resampled unit twice as in Step 2 above. Using a BOMA in Steps 1 and 2 of the procedure above also results in an approximately unbiased variance estimator. BOMA's are fairly easily constructed for even n_h using balanced incomplete block designs and Hadamard matrices, but are difficult to construct for odd n_h. They can also handle unequal n_h's for different strata, though construction becomes a more serious problem (see Sitter 1993).

6. A SIMULATION

To study the properties of the proposed resampling variance estimators, we consider a finite population of $H = 32$ strata with N_h clusters in stratum h and ten ultimate units in each cluster. The characteristic of interest y_{hik} are generated as follows:

Table 1
A Set of Balanced Resamples Constructed from a BOMA

	h						
b	1	2	3	4	5	6	7
1	(1,1,3,3)	(1,1,3,3)	(1,1,3,3)	(1,1,3,3)	(1,1,3,3)	(1,1,3,3)	(1,1,3,3)
2	(1,1,4,4)	(1,1,4,4)	(1,1,4,4)	(1,1,4,4)	(1,1,4,4)	(1,1,4,4)	(1,1,4,4)
3	(1,1,2,2)	(1,1,2,2)	(1,1,2,2)	(1,1,2,2)	(1,1,2,2)	(1,1,2,2)	(1,1,2,2)
4	(2,2,4,4)	(1,1,3,3)	(2,2,4,4)	(1,1,3,3)	(2,2,4,4)	(1,1,3,3)	(2,2,4,4)
5	(2,2,3,3)	(1,1,4,4)	(2,2,3,3)	(1,1,4,4)	(2,2,3,3)	(1,1,4,4)	(2,2,3,3)
6	(3,3,4,4)	(1,1,2,2)	(3,3,4,4)	(1,1,2,2)	(3,3,4,4)	(1,1,2,2)	(3,3,4,4)
7	(2,2,4,4)	(2,2,4,4)	(1,1,3,3)	(1,1,3,3)	(2,2,4,4)	(2,2,4,4)	(1,1,3,3)
8	(2,2,3,3)	(2,2,3,3)	(1,1,4,4)	(1,1,4,4)	(2,2,3,3)	(2,2,3,3)	(1,1,4,4)
9	(3,3,4,4)	(3,3,4,4)	(1,1,2,2)	(1,1,2,2)	(3,3,4,4)	(3,3,4,4)	(1,1,2,2)
10	(1,1,3,3)	(2,2,4,4)	(2,2,4,4)	(1,1,3,3)	(1,1,3,3)	(2,2,4,4)	(2,2,4,4)
11	(1,1,4,4)	(2,2,3,3)	(2,2,3,3)	(1,1,4,4)	(1,1,4,4)	(2,2,3,3)	(2,2,3,3)
12	(1,1,2,2)	(3,3,4,4)	(3,3,4,4)	(1,1,2,2)	(1,1,2,2)	(3,3,4,4)	(3,3,4,4)
13	(1,1,3,3)	(1,1,3,3)	(1,1,3,3)	(2,2,4,4)	(2,2,4,4)	(2,2,4,4)	(2,2,4,4)
14	(1,1,4,4)	(1,1,4,4)	(1,1,4,4)	(2,2,3,3)	(2,2,3,3)	(2,2,3,3)	(2,2,3,3)
15	(1,1,2,2)	(1,1,2,2)	(1,1,2,2)	(3,3,4,4)	(3,3,4,4)	(3,3,4,4)	(3,3,4,4)
16	(2,2,4,4)	(1,1,3,3)	(2,2,4,4)	(2,2,4,4)	(1,1,3,3)	(2,2,4,4)	(1,1,3,3)
17	(2,2,3,3)	(1,1,4,4)	(2,2,3,3)	(2,2,3,3)	(1,1,4,4)	(2,2,3,3)	(1,1,4,4)
18	(3,3,4,4)	(1,1,2,2)	(3,3,4,4)	(3,3,4,4)	(1,1,2,2)	(3,3,4,4)	(1,1,2,2)
19	(2,2,4,4)	(2,2,4,4)	(1,1,3,3)	(2,2,4,4)	(1,1,3,3)	(1,1,3,3)	(2,2,4,4)
20	(2,2,3,3)	(2,2,3,3)	(1,1,4,4)	(2,2,3,3)	(1,1,4,4)	(1,1,4,4)	(2,2,3,3)
21	(3,3,4,4)	(3,3,4,4)	(1,1,2,2)	(3,3,4,4)	(1,1,2,2)	(1,1,2,2)	(3,3,4,4)
22	(1,1,3,3)	(2,2,4,4)	(2,2,4,4)	(2,2,4,4)	(2,2,4,4)	(1,1,3,3)	(1,1,3,3)
23	(1,1,4,4)	(2,2,3,3)	(2,2,3,3)	(2,2,3,3)	(2,2,3,3)	(1,1,4,4)	(1,1,4,4)
24	(1,1,2,2)	(3,3,4,4)	(3,3,4,4)	(3,3,4,4)	(3,3,4,4)	(1,1,2,2)	(1,1,2,2)

$$y_{hik} = y_{hi} + \varepsilon_{hik},$$

where $y_{hi} \sim N(\mu_h, \sigma_h^2)$ independent of $\varepsilon_{hik} \sim N(0, [1-\rho]\sigma_h^2/\rho)$ and the parameter values are those given in Table 2. For a particular value of the intracluster correlation, ρ, a single finite population was thus generated and then fixed and repeatedly sampled from. Each simulation consisted of selecting $n_h = 2$ clusters with replacement from stratum h for $h = 1, \ldots, H$ and enumerating the entire cluster. Each ultimate unit in the obtained cluster was independently declared a respondent or nonrespondent with probability p and $(1-p)$ respectively, *i.e.*, uniform response. The nonrespondents were then imputed both using random imputation and adjusted random imputation and the population total and distribution function, for various values of $F(t)$, were estimated. Two values of ρ, 0.1 and 0.3, and two values of p, 0.6 and 0.8, were considered. Note that the first-stage sampling fraction is quite small (0.064), so that with-replacement and without-replacement sampling are essentially equivalent.

To compare the performance of the different variance estimators we calculated the percent relative bias and relative instability for each, defined as

$$\%\mathrm{RB} = \frac{100}{S}\sum_{s=1}^{S} v_s(\hat{\theta}_I)/\mathrm{MSE}(\hat{\theta}_I)$$

and

$$\mathrm{RI} = \left\{\frac{1}{S}\sum_{s=1}^{S}[v_s(\hat{\theta}_I) - \mathrm{MSE}(\hat{\theta}_I)]^2\right\}^{1/2} \Big/ \mathrm{MSE}(\hat{\theta}_I),$$

respectively, where the number of simulation runs was $S = 5{,}000$ and the true $\mathrm{MSE}(\hat{\theta}_I)$ was obtained through an independent set of 50,000 simulation runs. The bootstrap variance estimators were each based on $B = 2{,}000$ bootstrap resamples. We obtain results for estimating the variance of $\hat{\theta}_I$ equal to the imputed total and the imputed distribution function using: (i) the repeated half-sample bootstrap with proper Monte Carlo approximation, v_B, as in equation (8) and with improper Monte Carlo approximation replacing $\bar{\theta}^*_{I(.)}$ with $\hat{\theta}_I$, denoted v_{B2}; and (ii) the proper repeated BRR, v_{BRR}, as in equation (9) and the improper repeated BRR replacing $\bar{\theta}_{I(.)}$ with $\hat{\theta}_I$, denoted v_{BRR2}.

Table 3 summarizes the results for percent relative bias using random imputation and adjusted random imputation. Note that adjusted random imputation is not presented for estimating the population total, Y, as adjusted random imputation removes the imputation variance from the estimator and thus simpler methods of variance estimation are available (Chen *et al.* 2000). It is clear from the high %RB for v_{B2} and v_{BRR2} that one must not replace $\bar{\theta}_{I(.)}$ and $\bar{\theta}^*_{I(.)}$ by $\hat{\theta}_I$ in the bootstrap or the BRR, respectively. It is also clear that both the repeated half- sample bootstrap and the repeated BRR variance estimators, v_B and v_{BRR} have negligible bias when properly applied.

Table 2
Parameters of the Finite Population

h	N_h	μ_h	σ_h	h	N_h	μ_h	σ_h
1	13	200	20.0	17	31	150	15.0
2	16	175	17.5	18	31	140	14.0
3	20	150	15.0	19	31	130	13.0
4	25	190	19.0	20	34	120	12.0
5	25	165	16.5	21	34	110	11.0
6	25	190	19.0	22	34	100	10.0
7	25	180	18.0	23	34	150	15.0
8	28	170	17.0	24	37	125	12.5
9	28	160	16.0	25	37	100	10.0
10	28	180	18.0	26	37	150	15.0
11	31	170	17.0	27	37	125	12.5
12	31	160	16.0	28	39	100	10.0
13	31	150	15.0	29	39	75	7.5
14	31	180	18.0	30	42	75	7.5
15	31	170	17.0	31	42	75	7.5
16	31	160	16.0	32	42	75	7.5

Given the results of Table 3, we consider relative instability, RI, only for υ_B and υ_{BRR}. We also restrict our presentation to $\rho = 0.3$ and $p = 0.6$ as the RI results were qualitatively the same in the other three cases. These results are given in Table 4. As one can see, though the differences are small, υ_B is slightly more stable than υ_{BRR}. This was generally the case for all values of ρ and p. We also included the adjusted jackknife of Rao and Shao (1992) and the adjusted BRR of Shao *et al.* (1998) in simulations for $\theta = \bar{Y}$ and υ_B again was uniformly more stable. For example, with $\rho = 0.3$ and $p = 0.6$ as in Table 4, RI for the adjusted jackknife and the adjusted BRR were both 0.27. This may be because the reimputation approach has an advantage in estimating the component of the variance due to the imputation against the adjustment approach, provided the resample size is large enough to eliminate Monte Carlo error as is the case in our simulations. But, when the number of reimputations is moderate (like in the BRR with reimputation or the bootstrap with B = 1,000), this advantage is not entirely realized.

Table 3
% RB for $\upsilon_B, \upsilon_{B2}, \upsilon_{BRR}$ and υ_{BRR2}

	Random imputation				Adjusted random imputation			
Estimand	υ_{BRR}	υ_{BRR2}	υ_B	υ_{B2}	υ_{BRR}	υ_{BRR2}	υ_B	υ_{B2}
				$\rho = 0.1$ and $p = 0.6$				
$\bar{Y}$	0.00	21.54	0.79	21.60				
F(t) = 0.0625	-1.09	15.92	-0.52	15.88	0.46	19.64	1.24	19.51
F(t) = 0.2500	-0.13	19.44	0.62	19.55	0.85	14.86	1.80	15.08
F(t) = 0.5000	-0.36	21.68	0.52	21.55	0.55	10.73	1.24	10.76
F(t) = 0.7500	-0.84	19.89	0.13	20.09	-0.36	10.98	0.54	11.31
F(t) = 0.9375	0.05	21.92	0.57	21.66	0.81	19.12	1.39	18.91
				$\rho = 0.1$ and $p = 0.8$				
$\bar{Y}$	-0.63	15.06	0.36	15.37				
F(t) = 0.0625	-1.99	10.30	-1.72	10.16	-1.65	10.97	-1.08	11.13
F(t) = 0.2500	-1.27	13.65	-0.88	13.30	-0.95	8.89	-0.52	8.81
F(t) = 0.5000	-0.72	15.26	0.02	15.26	-0.12	6.58	0.25	6.53
F(t) = 0.7500	-0.37	14.50	0.57	14.76	0.36	7.56	1.05	7.81
F(t) = 0.9375	-0.14	16.16	0.75	16.36	0.56	13.04	1.22	13.08
				$\rho = 0.3$ and $p = 0.6$				
$\bar{Y}$	0.25	21.34	0.78	21.09				
F(t) = 0.0625	-1.39	11.45	-0.86	11.37	-0.35	15.38	0.64	15.64
F(t) = 0.2500	-0.41	19.89	0.14	19.73	1.23	13.79	1.71	13.62
F(t) = 0.5000	-0.10	20.25	0.37	19.89	0.29	8.97	0.78	8.88
F(t) = 0.7500	-1.40	16.70	-0.49	16.89	-0.75	9.24	0.07	9.49
F(t) = 0.9375	0.71	17.78	1.03	17.57	0.91	15.07	1.34	15.04
				$\rho = 0.3$ and $p = 0.8$				
$\bar{Y}$	0.01	15.22	0.93	15.51				
F(t) = 0.0625	-1.09	7.54	-0.56	7.69	-1.24	8.64	-0.35	9.07
F(t) = 0.2500	-0.44	15.22	-0.08	14.99	-0.23	8.18	0.29	8.23
F(t) = 0.5000	0.05	14.92	0.71	14.84	0.43	6.21	0.86	6.20
F(t) = 0.7500	0.13	12.54	0.86	12.70	0.81	6.85	1.26	6.99
F(t) = 0.9375	1.62	13.13	2.06	13.01	1.86	11.04	2.34	11.02

Table 4
RI for v_B and v_{BRR} with $\rho = 0.3$ and $p = 0.6$

	Random imputation		Adjusted random imputation	
Estimand	v_{BRR}	v_B	v_{BRR}	v_B
Y	0.27	0.23		
$F(t) = 0.0625$	0.60	0.59	0.57	0.56
$F(t) = 0.2500$	0.35	0.32	0.37	0.35
$F(t) = 0.5000$	0.27	0.23	0.28	0.26
$F(t) = 0.7500$	0.29	0.26	0.30	0.28
$F(t) = 0.9375$	0.48	0.46	0.48	0.46

7. CONCLUSION

We proposed repeated half-sample bootstrap and balanced repeated replication methods for variance estimation in the presense of random imputation that capture the imputation variance by reimputing for each replication using the same random imputation method as in the original sample. These repeated half-sample methods are valid in stratified multi-stage sampling, even when the number of psu's sampled in each stratum is very small, *e.g.*, 2. The key is that these methods use a stratum resample size that is equal to the original sample size without resorting to rescaling. These provide a unified method that works irrespective of the imputation method (random or non-random), the stratum size (small or large), the type of estimator (smooth or nonsmooth), or the type of problem (variance estimation or sampling distribution estimation). It is important to note that using reimputation to capture the imputation variance requires that one take greater care in the definition of the BRR and the Monte Carlo approximation to the bootstrap variance. In both cases it is important to use the mean of the replicates in the definition as opposed to replacing it with the estimator applied to the original sample.

ACKNOWLEDGEMENTS

Hiroshi Saigo was supported by grants from the Promotion and Mutual Aid Corporation for Private Universities of Japan and the Japan Economic Research Foundation. Jun Shao was supported by National Science Foundation Grant DMS-0102223, and National Security Agency Grant MDA904-99-1-0032. Randy R. Sitter was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. The authors would like to thank all referees for their helpful comments and suggestions.

REFERENCES

CHEN, J., RAO, J.N.K. and SITTER, R.R. (2000). Adjusted imputation for missing data in complex surveys. *Statistica Sinica*, 10, 1153-1169.

EFRON, B. (1994). Missing data, imputation, and the bootstrap. *Journal of the American Statistical Association*, 89, 463-479.

KREWSKI, D., and RAO, J.N.K. (1981). Inference from stratified samples: properties of the linearization, jackknife, and balanced repeated replication methods. *The Annals of Statistics*, 9, 1010-1019.

MANTEL, H.J., and SINGH, A.C. (1991). Standard errors of estimates of low proportions: A proposed methodology. Technical Report, Statistics Canada.

MCCARTHY, P.J. (1969). Pseudoreplication half samples. *Review of the International Statistical Institute*, 37, 239-264.

NIGAM, A.K., and RAO, J.N.K. (1996). On balanced bootstrap, for stratified multistage samples. *Statistica Sinica*, 6, 199-214.

RAO, J.N.K. (1993). Linearization variance estimators under imputation for missing data. Technical Report, Laboratory for Research in Statistics and Probability, Carleton University.

RAO, J.N.K., and SHAO, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811-822.

RAO, J.N.K., and SITTER, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82, 453-460.

RAO, J.N.K., and WU, C.F.J. (1985). Inference from stratified samples: second order analysis of three methods for non-linear statistics. *Journal of the American Statistical Association*, 80, 620-630.

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

RUBIN, D.B. (1978). Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 20-34.

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

RUBIN, D.B., and SCHENKER, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81, 366-374.

SHAO, J., CHEN, Y. and CHEN, Y. (1998). Balanced repeated replications for stratified multistage survey data under imputation. *Journal of the American Statistical Association*, 93, 819-831.

SHAO, J., and RAO, J.N.K. (1994). Standard errors for low income proportions estimated from stratified multi-stage samples. *Sankhyā, B*, Special Volume 55, 393-414.

SHAO, J., and SITTER, R.R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association*, 91, 1278-1288.

SHAO, J., and WU, C.F.J. (1992). Asymptotic properties of the balanced repeated replication method for sample quantiles. *The Annals of Statistics*, 20, 1571-1593.

SITTER, R.R. (1993). Balanced repeated replications based on orthogonal multi-arrays. *Biometrika*, 80, 211-221.

SITTER, R.R. (1997). Variance estimation for the regression estimator in two-phase sampling. *Journal of the American Statistical Association*, 92, 780-787.

Survey Methodology, December 2001
Vol. 27, No. 2, pp. 197-203
Statistics Canada

Local Polynomial Regression in Complex Surveys

D.R. BELLHOUSE and J.E. STAFFORD[1]

ABSTRACT

Local polynomial regression methods are put forward to aid in exploratory data analysis for large-scale surveys. The proposed method relies on binning the data on the x-variable and calculating the appropriate survey estimates for the mean of the y-values at each bin. When binning on x has been carried out to the precision of the recorded data, the method is the same as applying the survey weights to the standard criterion for obtaining local polynomial regression estimates. The alternative of using classical polynomial regression is also considered and a criterion is proposed to decide whether the nonparametric approach to modeling should be preferred over the classical approach. Illustrative examples are given from the 1990 Ontario Health Survey.

KEY WORDS: Covariates; Exploratory data analysis; Kernel smoothing; Regression.

1. INTRODUCTION

Following Fuller (1975), multiple linear regression techniques have been studied and used extensively in sample surveys. At least three chapters of Skinner, Holt and Smith (1989) are devoted to this subject. Here we restrict attention to the case in which there is one covariate x for the variate of interest y so that we could consider polynomial regression as well as simple linear regression. In this context we could also consider the nonparametric approach of local polynomial regression, which, for the case of independent and identically distributed random variables, is described in Hardle (1990), Wand and Jones (1995), Fan and Gijbels (1996), Simonoff (1996) and Eubank (1999). Using the survey weights, Korn and Graubard (1998) introduced the use of local polynomial regression for graphical display of complex survey data. However, they did not provide any statistical properties for their procedures. Smith and Njenga (1992) used regression kernel smoothing techniques to obtain robust estimates of the mean and regression parameters for an assumed superpopulation model. Here we use local polynomial regression as an exploratory tool to discover relationships between y and its covariate x.

We assume that the covariate x is measured on a continuous scale. Due to the precision at which the data are recorded for the survey file and the size of the sample, there will be multiple observations at many of the distinct values. This feature of large-scale survey data has been exploited by Hartley and Rao (1968, 1969) in their scale-load approach to the estimation of finite population parameters. Here we exploit this same feature of the data to examine the relationship between y and its covariate x. In recognizing that the data may be naturally binned to the precision of the data, we can consider taking a further step by constructing larger bin sizes. Under this approach we examine the effect of the sampling design on estimates and second order moments.

Suppose that in the finite population of size N, x has k distinct values so that natural binning has taken place, or that x has been categorized into k bins that are wider than the precision of the data. Let x_i be the value of x representing the i^{th} bin, and assume that the values of x_i are equally spaced. The spacing or bin size $b = x_i - x_{i-1}$. The finite population mean for the y-values at x_i is $\bar{y}_i$. We assume that a sample of size n taken from this population has the same structure as the population in that there are k bins. From the sample data we calculate the survey estimate of $\hat{\bar{y}}_i$ of $\bar{y}_i$. The finite population proportion of the observations with value x_i is denoted by p_i. This proportion is estimated by the survey estimate $\hat{p}_i$. We assume that $\hat{\bar{y}}_i$ and $\hat{p}_i$ are asymptotically unbiased, in the sense of Särndal, Swensson and Wretman (1992, pages 166-167), for $\bar{y}_i$ and p_i respectively. The survey estimates $\hat{\bar{y}}_i$ for $i = 1, ..., k$ have variance-covariance matrix $\mathbf{V}$. On considering the distinct values x_i as domains, the estimated variance-covariance matrix $\hat{\mathbf{V}}$ may be obtained easily through survey packages such as SUDAAN and STATA.

There are several advantages to binning the data on the covariate x for exploratory data analysis:

- For large surveys, a plot of $\hat{\bar{y}}_i$ against x_i may be more informative and less cluttered than a plot of the raw data.
- By appealing to a finite population central limit theorem on $\hat{\bar{y}}_i$ and imposing a superpopulation assumption on $\bar{y}_i$, a relatively simple model for $\hat{\bar{y}}_i$ may be assumed so that the analyst may easily focus on the central issue considered here, determination of the trend function in x.

[1] D.R. Bellhouse Department of Statistical and Actuarial Sciences, Western Science Centre, University of Western Ontario, London, Ontario N6A 5B7, e-mail: bellhouse@stats.uwo.ca; J.E. Stafford, Department of Public Health Sciences, Faculty of Medecine, McMurrich Building, University of Toronto, Toronto, Ontario, M5S 1A8, e-mail: stafford@utstat.toronto.edu.

- Once $\hat{\mathbf{V}}$ has been obtained, then a wide variety of powerful exploratory data analyses can be easily carried out in languages such as S-Plus. Working with the raw data requires continued appeals to SUDAAN or STATA for the appropriate variance estimates.
- By binning the data, an approach to regression analysis is obtained that provides a parallel to other nonparametric approaches to survey data analysis. For example, in categorical data analysis obtained initially by Rao and Scott (1981), in the logistic regression approach of Roberts, Rao and Kumar (1987) or in the generalized linear model approach of Bellhouse and Rao (2000), the tests and associated distributions are obtained through survey estimates of domain means or proportions.

For the superpopulation, we assume that we have a model such that $E_m(\bar{y}_i) = m(x_i)$, where E_m is the superpopulation expectation. We assume further that as we move to a continuum of values on x, then $m(x)$ is a smooth function. The function $m(x)$ is the ultimate function of interest for estimation. In section 2 we provide local polynomial regression methods to estimate $m(x)$. These methods are applied to data from the 1990 Ontario Health Survey in section 3. In section 4, the question is asked: would the classical polynomial regression techniques have served equally as well in modeling $m(x)$? Some future directions for this work are given in section 5. Generally, we adopt the notation of Wand and Jones (1995) in discussing local polynomial regression here.

2. BASIC METHODOLOGY

For local polynomial regression, the nestimate of $m(x)$ at any value of x is obtained upon minimizing

$$\sum_{i=1}^{k} \hat{p}_i \left\{ \hat{\bar{y}}_i - \beta_0 - \beta_1 (x_i - x) - \ldots - \beta_q (x_i - x)^q \right\}^2 K\left((x_i - x)/h\right)/h \qquad (1)$$

with respect to $\beta_0, \beta_1, \ldots, \beta_q$. The values that minimize (1) are denoted by $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_q$. Further, for the given value of x, $\hat{m}(x) = \hat{\beta}_0$. In (1), the kernel $K(t)$ is a symmetric function with $\int K(t)\,dt = 1$, $\int tK(t)\,dt = 0$, $0 < \int t^2 K(t)\,dt < \infty$ and

$$R(K) = \int [K(t)]^2\,dt < \infty. \qquad (2)$$

Also in (1), h is the window width of the kernel. In minimizing (1) to obtain local polynomial regression estimates, there are two possibilities for binning on x. The first is to bin to the precision of the recorded data so that $\hat{\bar{y}}_i$ is calculated at each distinct outcome of x. In other situations it may be practical to pursue a binning on x that is rougher than the accuracy of the data.

In moving from the sample to the population we maintain the same window width h. This is in contrast to Breidt and Opsomer (2000) and Buskirk (1999) who assume a smoothing parameter h_N for smoothing in the full finite population. In the context here, this would yield a function $m_N(x)$, the finite population smoothed version of the $\bar{y}_i$ with smoothing parameter h_N, as a finite population parameter of interest followed by $m(x)$ the hypothetical smooth function under the asymptotic assumptions. We have kept h constant in view of the way in which binning that has been done; the bin structure is the same in the sample as in the population. The choice of the smoothing parameter h depends on the spacing of the x's and the variation in the data (Green and Silverman 1994, pages 43-44). The spacing of the covariate is usually dominant in the determination of h. Since the spacing has been kept constant from sample to finite population with the spacing changing only when the asymptotic assumptions are applied, we keep $h_N = h$.

Korn and Graubard (1998) provide a slightly different objective function to (1). They replace the sum over the bins in (1) by the sum over all sampled units and $\hat{p}_i$ in (1) by the sample weights. Korn and Graubard's objective function reduces to (1) plus a term that involves the weighted sum of squares of deviations of sample observations from the binned means where the weights are the sample weights scaled to sum to one. Consequently, the estimate of $m(x)$ is the same in both cases.

The estimate $\hat{m}(x)$ and its first two moments can be expressed in matrix notation. The forms are exactly the same as those that appear, for example, in Wand and Jones (1995, chapter 5.3) whose notation we have adopted. Let the vector of finite population means at the distinct values of x be $\bar{\mathbf{y}} = (\bar{y}_1, \ldots, \bar{y}_k)^{\mathrm{T}}$ and let $\hat{\bar{\mathbf{y}}}$ be its vector of survey estimates. Further, let

$$\mathbf{X}_x = \begin{bmatrix} 1 & x_1 - x & \cdots & (x_1 - x)^q \\ 1 & x_2 - x & \cdots & (x_2 - x)^q \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_k - x & \cdots & (x_k - x)^q \end{bmatrix}$$

and

$$\mathbf{W}_x = \frac{1}{h} \operatorname{diag}\left(p_1 K\left((x_1 - x)/h\right), p_2 K\left((x_2 - x)/h\right), \cdots p_k K\left((x_k - x)/h\right)\right).$$

The matrix $\hat{\mathbf{W}}_x$ is $\mathbf{W}_x$ with p replaced by $\hat{p}$. Then

$$\hat{m}(x) = \mathbf{e}^{\mathrm{T}} (\mathbf{X}_x^{\mathrm{T}} \hat{\mathbf{W}}_x \mathbf{X}_x)^{-1} \mathbf{X}_x^{\mathrm{T}} \hat{\mathbf{W}}_x \hat{\bar{\mathbf{y}}}, \qquad (3)$$

where $\mathbf{e}$ is the $k \times 1$ vector $(1, 0, 0, .., 0)^{\mathrm{T}}$. The approximate design-based expectation of $\hat{m}(x)$ is

$$E_p(\hat{m}(x)) = \mathbf{e}^{\mathrm{T}} (\mathbf{X}_x^{\mathrm{T}} \mathbf{W}_x \mathbf{X}_x)^{-1} \mathbf{X}_x^{\mathrm{T}} \mathbf{W}_x \bar{\mathbf{y}}, \qquad (4)$$

where E_p denotes expectation with respect to the sampling design. We can also consider (4) as a smoothed estimate of $m(x)$ so that $\hat{m}(x)$ is also an estimate of $m(x)$. In the derivation of (4) we note that $E_p(\hat{\bar{y}}) = \bar{y}$ and $E_p(\hat{\mathbf{W}}_x) = \mathbf{W}_x$ for large sample size n. Further, in (3) we can write $\hat{\mathbf{W}}_x = \mathbf{W}_x + \hat{\mathbf{A}}$, where $\hat{\mathbf{A}} = \hat{\mathbf{W}}_x - \mathbf{W}_x$. We use the first two terms in the expansion $(\mathbf{I} + \mathbf{B})^{-1} = \mathbf{I} - \mathbf{B} + \mathbf{B}^2 - \mathbf{B}^3 + \cdots$ as an approximation to complete the derivation. Using the same techniques, the approximate design-based variance is given by

$$V_p(\hat{m}(x)) = \mathbf{e}^T (\mathbf{X}_x^T \mathbf{W}_x \mathbf{X}_x)^{-1} \mathbf{X}_x^T \mathbf{W}_x \mathbf{V} \mathbf{W}_x \mathbf{X}_x (\mathbf{X}_x^T \mathbf{W}_x \mathbf{X}_x)^{-1} \mathbf{e}. \quad (5)$$

The results in (4) and (5) were obtained ignoring higher order terms in $1/n$. An estimate of the variance $\hat{V}_p(\hat{m}(x))$ is obtained on substituting the survey estimate $\hat{\mathbf{V}}$ for $\mathbf{V}$ and $\hat{\mathbf{W}}_x$ for $\mathbf{W}_x$ in (5).

3. EXAMPLES FROM THE ONTARIO HEALTH SURVEY

We illustrate local polynomial regression techniques with data from the Ontario Health Survey (Ontario Ministry of Health 1992). This survey was carried out in 1990 using a stratified two-stage cluster sample. The purpose was to measure the health status of the people of Ontario and to collect data relating to the risk factors of major causes of morbidity and mortality in Ontario. The survey was designed to be compatible with the Canada Health Survey carried out in 1978-79. A total sample size of 61,239 people was obtained from 43 public health units across Ontario. The public health unit was the basic stratum with an additional division of the health unit into rural and urban strata so that there were a total of 86 strata. The first stage units within a stratum were enumeration areas taken from the 1986 Census of Canada. An average of 46 enumeration areas was chosen within each stratum. Within an enumeration area, dwellings were selected, approximately 15 from an urban enumeration area and 20 from a rural enumeration area. Information was collected on members of the household within the dwelling.

Several health characteristics were measured. We focus on one continuous variable from the survey, Body Mass Index (BMI). The BMI is a measure of weight status and is calculated from the weight in kilograms divided by the square of the height in meters. The index is not applicable to adolescents, adults over 65 years of age and pregnant or breastfeeding women. The measure varies between 7.0 and 45.0. A value of the BMI less than 20.0 is often associated with health problems such as eating disorders. An index value above 27.0 is associated with health problems such as hypertension and coronary heart disease. Associated with the BMI is another measure, the Desired Body Mass Index (DBMI). The DBMI is the same measure as BMI with actual weight replaced by desired weight. A total of 44,457 responses were obtained for the BMI and 41,939 for the DBMI.

When there are only a few distinct outcomes of x, binning on x is done in a natural way. For example, in investigating the relationship between the body mass index (BMI) and age, the age of the respondent was reported only at integral values. The solid dots in Figure 1 are the survey domain estimates of the average BMI ($\hat{\bar{y}}_i$) for women at each of the ages 18 through 65 (x_i). The solid and dotted lines show the plot of $\hat{m}(x)$ against x using bandwidths $h = 7$ and $h = 14$ respectively. It may be seen from Figure 1 that BMI increases approximately linearly with age until around age 50. The increase slows in the early 50s, peaks at age 55 or so, and then begins to decrease. On plotting the trend lines only for BMI and the desired body mass index (DBMI) for females as shown in Figure 2, it may be seen that, on average, women desire to reduce their BMI at every age by approximately two units.

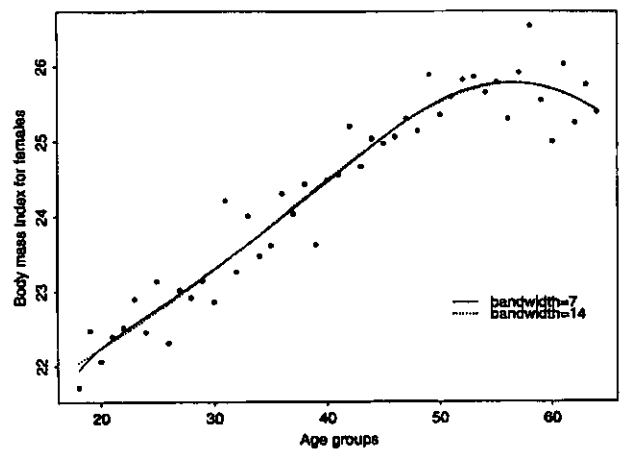

Figure 1. Age trend in BMI for females

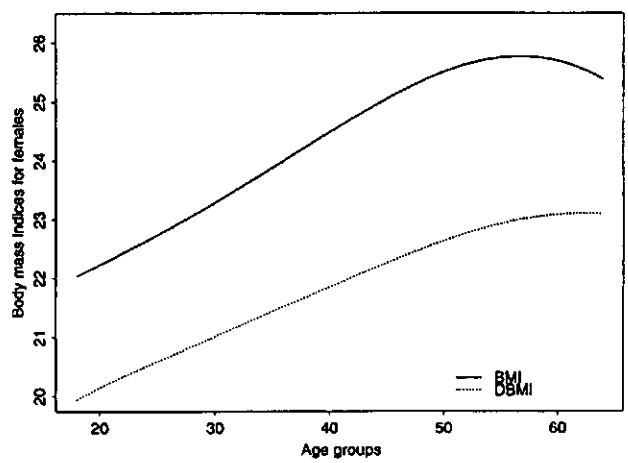

Figure 2. Age trends for females

In other situations it is practical to construct bins on x wider than the precision of the data. To investigate the relationship between what women desire for their weight (DBMI = $\hat{\bar{y}}_i$) and what women actually weigh (BMI = x_i) the x-values were grouped. Since the data were very sparse for values of BMI below 15 and above 42, these data were removed from consideration. The remaining groups were 15.0 to 15.2, 15.3 to 15.4 and so on, with the value of x_i chosen as the middle value in each group. The binning was done in this way for the purposes of illustration to obtain a wide range of equally spaced nonempty bins. For each group the survey estimate $\hat{\bar{y}}_i$ was calculated. The solid dots in Figure 3 show the survey estimates of women's DBMI for each grouped value of their respective BMI. The scatter at either end of the line reflects the sampling variability due to low sample sizes. The plot shows a slight desire to gain weight when the BMI is at 15. This desire is reversed by the time the BMI reaches 20 and the gap between the desire (DBMI) and reality (BMI) widens as BMI increases.

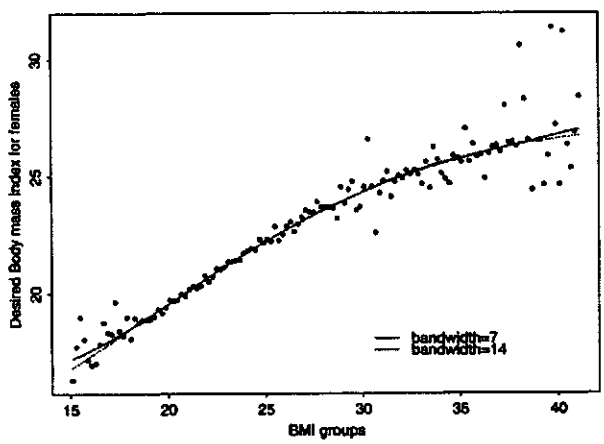

Figure 3. BMI trend in DBMI for females

4. PARAMETRIC VERSUS NONPARAMETRIC REGRESSION

Local polynomial regression allows us to obtain nonparametrically a functional relation between y and x. However, a parametric model may also be reasonable. For example, on examining Figure 1 showing the Body Mass Index against age, we might consider the parametric model that y has a quadratic relationship to x. We may also want to test in Figure 2 if the two lines are parallel, or equivalently that the difference between the Body Mass Index and the Desired Body Mass Index for females is constant over all ages. This would involve modeling the trend lines as second degree polynomials and testing for equality in the trend lines of the parameters associated with the quadratic term as well as the parameters associated with the linear term. In all cases, the question arises as to whether or not the data can be adequately modeled by a polynomial relationship between y and x. One method that we propose as an answer to this question is to calculate the confidence bands based on local polynomial regression. These bands can be thought of as providing a region of acceptable model representations. If an appropriate parametric regression line falls within the bands, then it provides a reasonable model description of the data. The $100(1-\alpha)\%$ local polynomial regression bands are obtained by ploting

$$\hat{m}(x) \pm z_{\alpha/2} \sqrt{\hat{V}_p(\hat{m}(x))} \tag{6}$$

over a range of values of x, where $z_{\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the standard normal distribution, where $\hat{m}(x)$ is determined from (3) and where $\hat{V}_p(\hat{m}(x))$ is (5) with $\mathbf{V}$ replaced by its sample estimate $\hat{\mathbf{V}}$.

The parametric regression line to be tested may be obtained in one of two ways depending upon what sample information is available. If the complete sample file with sampling weights is available, then the standard regression approach in, for example, SUDAAN may be used. If only the binned data are available, in particular the survey estimates $\hat{\bar{y}}_i$ with estimated variance-covariance matrix $\hat{\mathbf{V}}$, then another approach is needed.

For this second approach assume that $m(x_i) = \mathbf{x}_i^T \boldsymbol{\beta}$, where $\mathbf{x}_i^T = (1, x_i, x_i^2, ..., x_i^q)$ and where $\boldsymbol{\beta}^T = (\beta_0, \beta_1, ..., \beta_q)$ is the vector of regression coefficients. For the finite population we assume that $\bar{y}_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i$, where the errors are deviations of the actual finite from the model. For simplicity, we assume that these errors have mean 0 and variance-covariance matrix $\sigma^2\mathbf{I}$. Since the data are given by the survey estimates $\hat{\bar{y}}_i$ with variance-covariance matrix $\mathbf{V}$, the operative model is

$$\hat{\bar{y}}_i = \mathbf{x}_i^T \boldsymbol{\beta} + \delta_i, \tag{7}$$

where the δ_i have mean 0 and variance-covariance matrix $\sum = \sigma^2\mathbf{I} + \mathbf{V}$. The usual weighted least squares estimate of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \sum^{-1} \mathbf{X})^{-1} \mathbf{X}^T \sum^{-1} \hat{\bar{\mathbf{y}}}_i, \tag{8}$$

where the i^{th} row of $\mathbf{X}$ is $\mathbf{x}_i^T$, $i = 1, ..., k$. In terms of data analysis it is necessary to replace $\sum$ in (8) by its estimate $\hat{\sum}$. Now the survey estimate of $\mathbf{V}$ is $\hat{\mathbf{V}}$ so that it remains to find an estimate of σ^2. This may be obtained through rss = $(\hat{\bar{\mathbf{y}}} - \mathbf{X}\hat{\boldsymbol{\beta}})^T (\hat{\bar{\mathbf{y}}} - \mathbf{X}\hat{\boldsymbol{\beta}})$, the residual sum of squares, by one of two ways.

The first method is to approximate the expected residual sum of squares under model (7) and solve directly for σ^2. Upon using the expansion $(\mathbf{I} + \mathbf{B})^{-1} = \mathbf{I} - \mathbf{B} + \mathbf{B}^2 - \mathbf{B}^3 + ...$ we find

$$E(\text{rss}) \cong (n - q - 1)\sigma^2 + \text{tr}(\mathbf{V}) - \text{tr}(\mathbf{X}^T\mathbf{V}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}). \tag{9}$$

The estimate of σ^2 is obtained on setting *rss* equal to the right hand side of (8) with $\mathbf{V}$ replaced by $\hat{\mathbf{V}}$ and then solving for σ^2. This leads to an iterative approach to model fitting. An initial estimate of $\boldsymbol{\beta}$ is obtained from (8) with $\mathbf{V}$ replaced by the survey estimate $\hat{\mathbf{V}}$. Then σ^2 is estimated through (9) and a new estimate of $\boldsymbol{\beta}$ using $\hat{\sum} = \hat{\sigma}^2\mathbf{I} + \hat{\mathbf{V}}$ is obtained. The process is repeated until convergence is

obtained in the estimate of σ^2. If the estimate of σ^2 is negative, it is set to 0. The second method for estimating σ^2 is obtaining by first treating the errors in (7) as multivariate normal variables. Then a profile likelihood for σ^2 can be obtained on replacing β and $\mathbf{V}$ by their estimates. The most influential term in this profile likelihood is

$$\mathbf{r}^T(\sigma^2\mathbf{I} + \hat{\mathbf{V}})^{-1}\mathbf{r}, \qquad (10)$$

where $\mathbf{r} = \hat{\bar{\mathbf{y}}} - \mathbf{X}(\mathbf{X}^T(\sigma^2\mathbf{I} + \hat{\mathbf{V}})^{-1}\mathbf{X})^{-1}\mathbf{X}^T(\sigma^2\mathbf{I} + \hat{\mathbf{V}})^{-1}\hat{\bar{\mathbf{y}}}$ is the vector of residuals. An approximation to the profile likelihood estimate $\hat{\sigma}^2$ is that value of σ^2 which minimizes (10).

To provide examples of the question of the adequacy of parametric regression, we examined two different variables in the Ontario Health Survey and their relationship to the body mass index (BMI). These were age and fat consumption as a percentage of total energy consumption. For age the binning was natural and at the precision of the recorded data. Age was restricted to the range of 18 to 65 years since the index is not applicable outside this range and age was recorded in years. The scatterplot of BMI against age with the accompanying local polynomial regression line is shown in Figure 1. The survey data on fat consumption in percentages were recorded to three decimal places. Due to the sparseness of the data at the extremes we looked at fat consumption in the range of 14 to 56% of total energy consumption. Further, we binned the data on the covariate (fat consumption) using bins 14.0 up to 14.2, 14.2 up to 14.4 and so on; the midpoints of the bins (14.1, 14.3 and so on) were used as the x_i. At each bin the survey estimate $\hat{\bar{y}}_i$ for BMI was calculated. It is the binned data that appear as a scatterplot of BMI against fat consumption in Figure 5. The solid line in Figure 5 is the local polynomial regression line with $q = 1$ for BMI on fat content. As in Figure 3, the larger variability at the extremes reflects greater sampling variability due to smaller sample sizes at the extremes. From Figure 5 it appears that BMI increases slightly as fat consumption increases. Since the complete data file for the survey was available, regression lines for all variables were obtained through SUDAAN.

In Figure 4 the solid lines are the 95% confidence bands based on (6) and the dashed line is the parametric second degree polynomial regression line. Since the dashed line falls near the border for women in their thirties and outside the bands for women in their early sixties, a second degree polynomial barely adequately describes the relation between BMI and age. Another model might be preferable. Figure 6 shows the same 95% confidence bands but for the consumption of fat as a percentage of total energy consumption. In this case the dotted line is the simple linear regression line of BMI on fat consumption. For fat consumption the line falls completely within the confidence bands so that simple linear regression appears to be an adequate description of the model relationship.

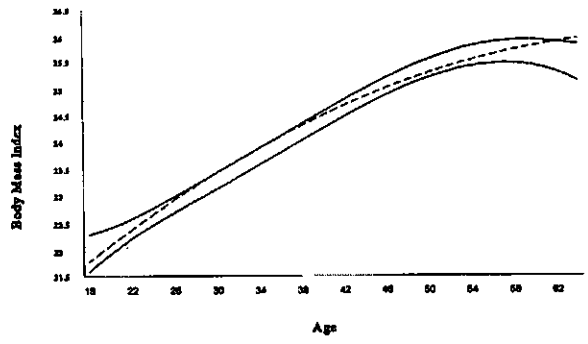

Figure 4. Confidence Bands for the Age Trend in BMI for Females

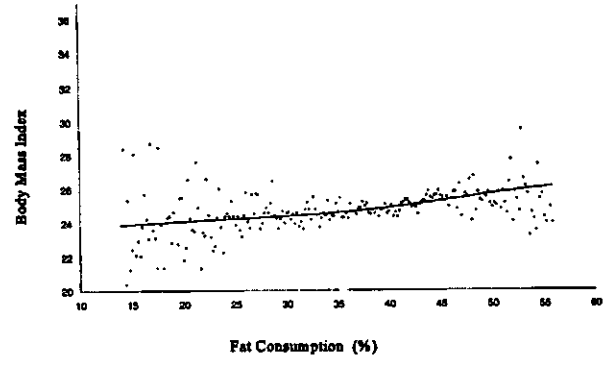

Figure 5. BMI Trend in Fat Consumption

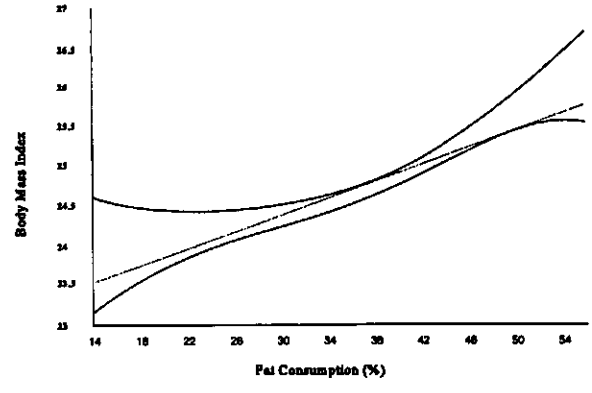

Figure 6. Confidence Bands for Fat Consumption Trend in BMI

If the data have been binned to the precision of the data as in the case of age above, and if the exploratory analysis is complete, we can stop. The estimates and variance estimates obtained are equal to the estimates and variance estimates obtained from the raw data. This may be seen on examining (3). The term on the right hand side of (3) can be expressed as a sum over the sample of the sample weights times a new measurement obtained from the raw y-measurement times an appropriate value taken from $\mathbf{e}^T(\mathbf{X}_x^T\hat{\mathbf{W}}_x\mathbf{X}_x)^{-1}\mathbf{X}_x^T\mathbf{W}_x^*$ times the total of the sample weights, where $\mathbf{W}_x^*$ is $\mathbf{W}_x$ with the p_i's removed. These

adjusted y-measurements may be fed into SUDAAN or STATA to obtain the required approximate variance estimate. It may be that the binning has been rougher than the precision of the data or that some bins have been dropped in the tails of the distribution of x due to sparseness of the data in those bins. Both of these situations occurred in analyzing the relationship of BMI to fat consumption. Once the exploratory analysis has been completed we can return with a final model and smoothing parameter, if a nonparametric approach is used in the final analysis, and apply to model to the raw data obtaining variance estimates through SUDAAN or STATA as necessary. Depending on the amount of roughness in the binning and the number of bins dropped due to sparseness in the data, the variance estimates obtained from the raw will be approximately the same as those from the binned data.

5. FUTURE DIRECTIONS

Like Bellhouse and Stafford (1999), this paper adapts a modern method of smoothing for the analysis of complex survey data. It represents an example of a host of regression techniques that could be used. To describe these we embed the current context in a general framework hinting at future work. In doing so we mimic the developments of Hastie and Tibshirani (1990).

Here a smoother is said to be linear if fitted values are obtained by applying a matrix $\mathbf{S}$ to a response vector $\mathbf{y}$. As in the case of simple linear regression for independent and identically distributed data, we let $\mathbf{H} = (\mathbf{X}^T\sum^{-1}\mathbf{X})^{-1}\mathbf{X}^T\sum^{-1}$ and further denote $(\mathbf{X}_x^T\hat{\mathbf{W}}_x\mathbf{X}_x)^{-1}\mathbf{X}_x^T\hat{\mathbf{W}}_x$ as $\mathbf{S}_p$. Both are examples of $\mathbf{S}$. In addition, the response vector of binned means is a type of smooth $\hat{\bar{y}}_i = \mathbf{S}_b\mathbf{y}$, where $\mathbf{y}$ is the vector of all sample responses and where $\mathbf{S}_b$ involves the sample weights. Also the usual regression context involves applying a matrix similar to $\mathbf{H}$ to the full response vector $\hat{\bar{y}}_i = \mathbf{H}_f\mathbf{y}$. So moving from usual regression to regressing means to local polynomial smoothing reduces to applying different smoothing matrices to $\mathbf{y}$:

$$\mathbf{H}_f\,\mathbf{y} \to \mathbf{H}\,\mathbf{S}_b\,\mathbf{y} \to \mathbf{S}_p\,\mathbf{S}_b\,\mathbf{y}.$$

In general $\mathbf{S}_p$ can be replaced by any smoother $\mathbf{S}$ and the methods extended to multiple covariates.

There are many advantages to binning the response from both a theoretical and practical standpoint. Standard smoothing tools, like those found in *Splus*, can be applied without modification of the smoother due to sampling issues. In addition, in the case of the additive model, finite population central limit theorems can be invoked and issues like degrees of freedom, choice of smoothing parameter, optimizing a criterion, can be handled in the usual manner. In the case of multiple covariates $x_1, ..., x_q$ the curse of dimensionality will result in sparse bins not allowing the use of the central limit theorem. This may be countered in the usual way by binning partial residuals one dimension at a time. Here smoothers $\mathbf{S}_j\mathbf{S}_{b_j}$, $j = 1, ..., q$ would be used in a backfitting algorithm.

It is our intention to study additive and generalized additive models in the above manner and to introduce these techniques to the analysis of complex survey data.

ACKNOWLEDGEMENTS

The authors would like to thank Rob Tibshirani for his helpful comments on this paper and the referees for their comments that helped to improve the presentation of the paper as well as to clarify some technical issues. This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

BELLHOUSE, D.R., and RAO, J.N.K. (2000). Analysis of domain means in complex surveys. *Journal of Statistical Planning and Inference*, to appear.

BELLHOUSE, D.R., and STAFFORD, J.E. (1999). Density estimation from complex surveys. *Statistica Sinica*, 9, 407-424.

BREIDT, F.J., and OPSOMER, J.D. (2000). Local polynomial regression estimators in survey sampling. Submitted for publication.

BUSKIRK, T. (1999). *Using Nonparametric Methods for Density Estimation with Complex Survey Data.* Ph.D. dissertation, Arizona State University.

EUBANK, R.L. (1999). *Nonparametric Regression and Spline Smoothing*. New York: Marcel Dekker.

FAN, J., and GIJBELS, I. (1996). *Local Polynomial Modelling and its Applications*. London: Chapman and Hall.

FULLER, W.A. (1975). Regression analysis for sample survey. *Sankhya* C, 37, 117-132.

GREEN, P.J., and SILVERMAN, B.W. (1994). *Nonparametric Regression and Generalized Linear Models*. London: Chapman and Hall.

HARDLE, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press: Cambridge.

HARTLEY, H.O., and RAO, J.N.K. (1968). A new estimation theory for sample surveys. *Biometrika*, 55, 547-557.

HARTLEY, H.O., and RAO, J.N.K. (1969). A new estimation theory for sample surveys, II. In *New Developments in Survey Sampling*, N.L. Johnson and H. Smith (Eds.) New York: Wiley Inter-Science, 147-169.

HASTIE, T.J., and TIBSHIRANI, R.J. (1990). *Generalized Additive Models*. London: Chapman and Hall.

KORN, E.L., and GRAUBARD, B.I. (1998). Scatterplots with survey data. *American Statistician*, 52, 58-69.

ONTARIO MINISTRY OF HEALTH (1992). *Ontario Health Survey: User's Guide, Volumes I and II.* Queen's Printer for Ontario.

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

ROBERTS, G., RAO, J.N.K. and KUMAR, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, 74, 1-12.

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

SIMONOFF, J.S. (1996). *Smoothing Methods in Statistics*. New York: Springer-Verlag.

SKINNER, C.J., HOLT, D. and SMITH, T.M.F. (1989). *Analysis of Complex Surveys*. New York: John Wiley and Sons.

SMITH, T.M.F., and NJENGA, E. (1992). Robust model-based methods for analytical surveys. *Survey Methodology*, 18, 187-208.

WAND, M.P., and JONES, M.C. (1995). *Kernel Smoothing*. London: Chapman and Hall.

Survey Methodology, December 2001
Vol. 27, No. 2, pp. 205-215
Statistics Canada

Modelling Compositional Time Series from Repeated Surveys

D.B.N. SILVA and T.M.F. SMITH[1]

ABSTRACT

A compositional time series is defined as a multivariate time series in which each of the series has values bounded between zero and one and the sum of the series equals one at each time point. Data with such characteristics are observed in repeated surveys when a survey variable has a multinomial response but interest lies in the proportion of units classified in each of its categories. In this case, the survey estimates are proportions of a whole subject to a unity-sum constraint. In this paper we employ a state space approach for modelling compositional time series from repeated surveys taking into account the sampling errors. The additive logistic transformation is used in order to guarantee predictions and signal estimates bounded between zero and one which satisfy the unity-sum constraint. The method is applied to compositional data from the Brazilian Labour Force Survey. Estimates of the vector of proportions and the unemployment rate are obtained. In addition, the structural components of the signal vector, such as the seasonals and the trends, are produced.

KEY WORDS: Additive logistic transformation; Compositional time series; Kalman Filter; Labour force survey; Repeated surveys; State space models.

1. INTRODUCTION

All surveys are multivariate and multipurpose, and most are longitudinal, repeating the same questions over time. There are two broad classes of repeated surveys, those with overlapping first stage units and those with no overlap of first stage units. Both designs admit a longitudinal macro-analysis of population aggregates but only the former allows a micro-analysis and the estimation of gross flows or some other similar unit level dynamic process. In this paper we explore the time series analysis of a multivariate vector of population aggregates, a macro-analysis, while taking into account the influence of the sampling errors of the survey using disaggregated data.

Denote by $\boldsymbol{\theta}_t = (\theta_{1t}, ..., \theta_{M+1,t})'$ a vector of population quantities of interest at time t, and assume that observations are made at equally spaced time intervals $t = 1, 2, ..., T$. Let $\boldsymbol{y}_t = (y_{1t}, ..., y_{M+1,t})'$ represent a survey-based estimate of $\boldsymbol{\theta}_t$ based on data collected at time t. Repeated surveys produce time series $\{\boldsymbol{y}_t\}$ comprising estimates of the unknown target series $\{\boldsymbol{\theta}_t\}$. Focussing on the unknown population vector $\boldsymbol{\theta}_t$, it is natural to imagine that knowledge of $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_{t-1}$ conveys useful information about $\boldsymbol{\theta}_t$ but without implying that it is perfectly predictable from $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_{t-1}$. One way of representing this situation is by considering $\boldsymbol{\theta}_t$ to be a random variable which evolves stochastically in time following a certain time series model, as first proposed for univariate survey analysis by Blight and Scott (1973), Scott and Smith (1974) and Scott, Smith and Jones (1977). The survey estimates $\boldsymbol{y}_t$ of $\boldsymbol{\theta}_t$ can then be written as:

$$\boldsymbol{y}_t = \boldsymbol{\theta}_t + \boldsymbol{e}_t \qquad (1)$$

where $\{\boldsymbol{\theta}_t\}$, $\{\boldsymbol{y}_t\}$ and $\{\boldsymbol{e}_t\}$ are random processes and $\boldsymbol{e}_t = (e_{1t}, ..., e_{M+1,t})'$ are the sampling errors such that $E(\boldsymbol{e}_t | \boldsymbol{\theta}_t) = \mathbf{0}$ and $V(\boldsymbol{e}_t | \boldsymbol{\theta}_t) = \sum_t$.

The early work of Scott *et al.* (1977) was concerned with univariate $\{y_t\}$ and distinguished different forms for the data available on $\{e_t\}$. If the only data available to the analyst are the population aggregate estimates $\{y_t\}$ then this is termed a secondary analysis and the examples in Scott *et al.* (1977) are based on a secondary analysis of survey data. If the individual data records are available, then variances and covariances can be estimated directly from the data and this is called a primary analysis. In addition, in the case of a rotating panel survey, elementary estimates (based on data from a set of units that join and leave the survey at the same time) can be used to estimate the covariance structure of the sampling errors. Subsequent work by Jones (1980) used a primary analysis to measure the structure of the sampling noise whereas Binder and Hidiroglou (1988), Binder and Dick (1989), Pfeffermann, Burck and Ben-Tuvia (1989), Pfeffermann and Burck (1990), Pfeffermann (1991), Binder, Bleuer and Dick (1993), Pfeffermann and Bleuer (1993), Pfeffermann, Bell and Signorelli (1996), Pfeffermann, Feder and Signorelli (1998) and Harvey and Chung (2000) employed an elementary analysis.

The time series analysis of survey data also requires that the signal process be modelled. In the early works it was assumed that $\{\theta_t\}$ was a stationary process and that $\{y_t\}$ was the superposition of two stationary processes therefore being itself stationary. Typically ARMA processes were assumed for $\{\theta_t\}$ and $\{e_t\}$, and hence for $\{y_t\}$. Binder and Hidiroglou (1988) wrote the processes in state space

[1] D.B.N. Silva, Instituto Brasileiro de Geografia e Estatística, Escola Nacional de Ciências Estatísticas, Rua André Cavalcanti 106 - Rio de Janeiro, RJ Brazil, 20231.050, e-mail: denisesilva@ibge.gov.br; T.M.F. Smith, University of Southampton, Faculty of Mathematical Studies, Highfield, Southampton, S017 1BJ, United Kingdom, e-mail: tmfs@maths.soton.ac.uk.

form which led rapidly to the introduction of nonstationary processes for the signal $\{\theta_t\}$, and structural models involving trends and seasonals have been used since then.

The aim is to improve estimation of the unobservable signal and its components, but when the sampling errors are autocorrelated these autocorrelations can induce spurious trends which get confounded with the true signal trend, as pointed out by Tiller (1992) and Pfeffermann, Bell and Signorelli (1996). When the variation in the sampling errors is not taken into account, their autocorrelation structure may be absorbed into either the seasonal or the trend components, thus affecting the inference from the model.

A special case of interest in repeated surveys is when the univariate target parameter $\{\theta_t\}$ is a proportion such as the unemployment rate. Unrestricted time series modelling of $\{\theta_t\}$ may lead to estimates outside the range $0 \le \theta_t \le 1$. Wallis (1987) used a logistic transformation to ensure that the estimates were bounded, however he failed to take into account the survey error. Pfeffermann (1991), Tiller (1992), Pfeffermann and Bleuer (1993), Pfeffermann, Bell and Signorelli (1996) fitted state space models to unemployment rate series taking into account survey errors but without using the logistic transformation to guarantee bounded estimates.

Most surveys are multivariate and there has been little work in the multivariate time series analysis of survey data. Brunsdon (1987) and Brunsdon and Smith (1998) analyse multivariate data from opinion polls taking into account the fact that the proportions are bounded and comprise a composition, but not allowing for the structure of the survey errors. This work provides useful insight into the modelling of time series of proportions. Compositional data have also been modelled using a state space approach, by Quintana and West (1988), Shephard and Harvey (1989) and Singh and Roberts (1992), but these authors also did not address the issue of modelling the autocovariance structure of the sampling errors when the observed compositions are obtained from repeated surveys.

The motivation for this work is that many variables investigated by statistical agencies have a multinomial response and interest lies in the estimation of the proportion of units classified in each of the categories. If this is the case, the vector of proportions sums to one and forms what is known as a composition. A compositional time series is therefore a multivariate time series comprising observations of compositions at each time point. We propose a class of multivariate state space models for compositional time series from repeated surveys, which takes into account the sampling errors and guarantees estimates satisfying the underlying constraints imposed by compositions. The procedure employs a signal-plus-noise structural model which yields seasonally adjusted series and estimates of the trend which satisfy the underlying sum constraint. The method is applied to compositional data from the Brazilian Labour Force Survey comprising estimates of the vector of proportions of labour market status. Estimates of seasonally adjusted compositions, trends and unemployment rate series are produced.

2. A FRAMEWORK FOR MODELLING COMPOSITIONAL DATA FROM OVERLAPPING SURVEYS

We assume that $\{\boldsymbol{\theta}_t\}$ is multivariate and the components θ_{mt} form a composition, *i.e.*, $0 < \theta_{mt} < 1 \ \forall \ m,t$ and $\sum_{m=1}^{M+1} \theta_{mt} = 1$. In this case $\boldsymbol{y}_t$ is a vector of sample estimates, based on the cross-sectional data of time t and belongs to the Simplex:

$$S^M = \{\boldsymbol{y}_t : 0 \le y_{mt} \le 1, \ m = 1,\ldots,M+1; \ \sum_{m=1}^{M+1} y_{mt} = 1; t = 1,\ldots,T\},$$

as in Brunsdon and Smith (1998). In addition, it is assumed that $\boldsymbol{y}_t$ is obtained from a survey with complex design and overlapping units between occasions. Since each of its components is subject to sampling errors, y_{mt} can be decomposed as:

$$y_{mt} = \theta_{mt} + e_{mt}, \quad m = 1,\ldots,M+1, \tag{2}$$

where θ_{mt} is the unknown population proportion assumed to follow a time series model, and e_{mt} is the sampling error. Considering the $M+1$ series simultaneously, (2) can be written in vector form as in equation 1. In addition, it is assumed that

$$\sum_{m=1}^{M+1} \theta_{mt} = \sum_{m=1}^{M+1} y_{mt} = 1 \quad \forall \ t, \tag{3}$$

which implies that $\sum_{m=1}^{M+1} e_{mt} = 0, \ \forall \ t$.

A compositional time series is a sequence of vectors $\boldsymbol{y}_t = (y_{1t},\ldots,y_{M+1,t})'$ each belonging to S^M. Aitchison (1986) examined the difficulties of applying standard methods to modelling and analysing compositions and suggested the use of transformations to map compositions from the Simplex S^M onto $\mathbb{R}^M$. One such transformation is the *additive logratio transformation* (a_M), defined in Aitchison (1986, page 113), which was first adopted in a time series context by Brunsdon (1987, page 75). The transformation is given by $\boldsymbol{v}_t = a_M(\boldsymbol{y}_t) = (v_{1t},\ldots,v_{Mt})'$, with

$$v_{mt} = \log\left(\frac{y_{mt}}{y_{M+1,t}}\right), \quad m = 1,\ldots,M, \quad \forall \ t, \tag{4}$$

where *log* denotes the natural logarithm. Note that $y_{M+1,t} = 1 - \sum_{m=1}^{M} y_{mt}$, sometimes called the fill-up value, is used as the reference variable or category. The inverse transformation, known as the *additive logistic transformation*, is given by $\boldsymbol{y}_t = a_M^{-1}(\boldsymbol{v}_t) = (y_{1t},\ldots,y_{M+1,t})'$ such that

$$y_{mt} = \begin{cases} \dfrac{\exp(v_{mt})}{1 + \sum_{j=1}^{M} \exp(v_{jt})} & m = 1,\ldots,M, \quad \forall\, t, \\ \dfrac{1}{1 + \sum_{j=1}^{M} \exp(v_{jt})} & m = M+1, \quad \forall\, t. \end{cases} \tag{5}$$

The state space modelling procedure for compositional time series is invariant to the choice of the reference variable (Silva 1996), and so any element $y_{mt} \neq y_{M+1,t}$ of $\boldsymbol{y}_t$ can be taken as the reference variable when applying the additive logistic transformation to the vector of survey estimates. When the logratios $\boldsymbol{v}_t$ are normally distributed the $M+1$ – part composition has an additive logistic normal distribution as defined in Aitchison and Shen (1980). For compositional time series, Brunsdon (1987) recommended the use of Vector ARMA models (Tiao and Box 1981) for the transformed series.

We propose a procedure that not only provides predictions and filtered estimates that are bounded between zero and one and satisfy the unity-sum constraint, but also improves the estimation of the unobservable signal and its components, taking into account the sampling error.

Following Bell and Hillmer (1990), the model in (2) can be rewritten as:

$$y_{mt} = \theta_{mt}\left(1 + \frac{e_{mt}}{\theta_{mt}}\right) = \theta_{mt}\, u_{mt}, \tag{6}$$

with

$$u_{mt} = \left(1 + \frac{e_{mt}}{\theta_{mt}}\right) = (1 + \tilde{u}_{mt}), \tag{7}$$

where $\tilde{u}_{mt} = e_{mt}/\theta_{mt}$ represents the relative sampling error of the estimated proportion.

Applying the additive logratio transformation defined in Aitchison (1986, page 113) to the vector $\boldsymbol{y}_t$, with components given in (2), produces a transformed vector $\boldsymbol{v}_t = a_M(\boldsymbol{y}_t) = (v_{1t}, \ldots, v_{Mt})'$ contained in $\mathbb{R}^M$. If $y_{M+1,t}$ is used as the reference variable, the transformed vector has as its m^{th} component:

$$v_{mt} = \log\left(\frac{y_{mt}}{y_{M+1,t}}\right) = \log\left(\frac{\theta_{mt}\, u_{mt}}{\theta_{M+1,t}\, u_{M+1,t}}\right)$$

$$= \log\left(\frac{\theta_{mt}}{\theta_{M+1,t}}\right) + \log\left(\frac{u_{mt}}{u_{M+1,t}}\right), \quad m = 1,\ldots,M. \tag{8}$$

From (8), a vector model for the transformed series can be written as:

$$\boldsymbol{v}_t = \boldsymbol{\theta}_t^* + \boldsymbol{e}_t^*, \tag{9}$$

with $\boldsymbol{v}_t = (v_{1t}, \ldots, v_{Mt})'$, $\boldsymbol{\theta}_t^* = (\theta_{1t}^*, \ldots, \theta_{Mt}^*)'$ and $\boldsymbol{e}_t^* = (e_{1t}^*, \ldots, e_{Mt}^*)'$, where $v_{mt} = \log(y_{mt}/y_{M+1,t})$, $\theta_{mt}^* = \log(\theta_{mt}/\theta_{M+1,t})$ and $e_{mt}^* = \log(u_{mt}/u_{M+1,t})$, for $m = 1, \ldots, M$. Note that model (9) has the same form as model (1).

To describe the survey data, model (9) must incorporate time series models for both $\{\boldsymbol{\theta}_t^*\}$ and $\{\boldsymbol{e}_t^*\}$. Hence a multivariate model for the transformed data will depend on the form of the time series models for $\{\boldsymbol{\theta}_t^*\}$ and $\{\boldsymbol{e}_t^*\}$.

The state space formulation for compositional data is examined in section 3, the model estimation is considered in section 4 and is illustrated using Brazilian Labour Force Survey data in section 5.

3. MODELLING THE TRANSFORMED SERIES

Our approach is based on assuming that the transformed series $\boldsymbol{v}_t = a_M(\boldsymbol{y}_t)$ has the signal plus noise structure in equation 9. We propose structural time series models for $\{\boldsymbol{\theta}_t^*\}$, as in Harvey (1989), and vector ARMA models (Tiao and Box 1981) for $\{\boldsymbol{e}_t^*\}$.

The transformed signal process $\{\boldsymbol{\theta}_t^*\}$ is assumed to follow the multivariate basic structural model, with each of the components $\{\theta_{mt}^*\}$ following a basic structural time series model (BSM) with possibly different parameters across the series. The cross-sectional relationship between the series is accounted for by the correlation structure of the system disturbances. The model for $\{\theta_{mt}^*\}$, $m = 1, 2, \ldots, M$, is then given by:

$$\begin{cases} \theta_{mt}^* = L_{mt}^* + S_{mt}^* + I_{mt}^*, \\ L_{mt}^* = L_{m,t-1}^* + R_{m,t-1}^* + \eta_{mt}^{(l)}, \\ R_{mt}^* = R_{m,t-1}^* + \eta_{mt}^{(r)}, \\ S_{mt}^* = -\sum_{j=1}^{11} S_{m,t-1}^* + \eta_{mt}^{(s)}, \end{cases} \tag{10}$$

where L_{mt}^* is the trend/level component of the signal θ_{mt}^*, R_{mt}^* is the corresponding change in the level, S_{mt}^* is the seasonal component and I_{mt}^* is an irregular component. For each component, the disturbances $\eta_{mt}^{(l)}$, $\eta_{mt}^{(r)}$, $\eta_{mt}^{(s)}$, and the irregular I_{mt}^*, are assumed to be mutually uncorrelated normal deviates with mean zero and variances $\sigma_{m_l}^2$, $\sigma_{m_r}^2$, $\sigma_{m_s}^2$, $\sigma_{m_I}^2$, respectively. That is, the $M \times 1$ vector disturbances $\boldsymbol{\eta}_t^{(l)}$, $\boldsymbol{\eta}_t^{(r)}$, $\boldsymbol{\eta}_t^{(s)}$ and $\boldsymbol{I}_t^*$, are mutually uncorrelated in all time periods. In addition, the irregulars I_{mt}^*, $I_{j(t-h)}^*$, with $m \neq j$, $h = \cdots, -2, -1, 0, 1, 2, \cdots$, are assumed to be correlated when $h = 0$, but uncorrelated for $h \neq 0$ and $\boldsymbol{I}_t^*$ has covariance matrix $\sum_I$. The same happens with the

system disturbances $\eta_{mt}^{(a)}$, $\eta_{j(t-h)}^{(a)}$, $a = l, r, s$, which are also correlated when $h = 0$, but uncorrelated for $h \neq 0$, with covariance matrices $\sum_l, \sum_r, \sum_s$. At each time t, the correlation structure between the components of the composition is summarized by $\sum_l$ and a block diagonal matrix with the blocks being $\sum_l, \sum_r, \sum_s$. Note that the relation between the series arises via the non-zero off-diagonal elements of the disturbance covariance matrices. The multivariate model (10) for $\{\boldsymbol{\theta}_t^*\}$ has the following state space formulation:

$$\begin{cases} \boldsymbol{\theta}_t^* &= \boldsymbol{H}^{(\theta)} \boldsymbol{\alpha}_t^{(\theta)} + \boldsymbol{I}_t^*; \\ \boldsymbol{\alpha}_t^{(\theta)} &= \boldsymbol{T}^{(\theta)} \boldsymbol{\alpha}_{t-1}^{(\theta)} + \boldsymbol{G}^{(\theta)} \boldsymbol{\eta}_t^{(\theta)}, \end{cases} \tag{11}$$

where $\boldsymbol{H}^{(\theta)} = [1010000000000] \otimes \boldsymbol{I}_M$,

$$\boldsymbol{\alpha}_t^{(\theta)} = [L_{1t}^* \ldots L_{Mt}^* R_{1t}^* \ldots R_{Mt}^* S_{1t}^* \ldots S_{Mt}^* \ldots S_{1,t-10}^* \ldots S_{M,t-10}^*]'$$

$$\boldsymbol{\eta}_t^{(\theta)} = (\eta_{1t}^{(l)} \ldots \eta_{Mt}^{(l)} \eta_{1t}^{(r)} \ldots \eta_{Mt}^{(r)} \eta_{1t}^{(s)} \ldots \eta_{Mt}^{(s)})',$$

$$\boldsymbol{G}^{(\theta)} = \begin{bmatrix} \boldsymbol{I}_3 \\ \ldots \ldots \\ \boldsymbol{0}_{10\times 3} \end{bmatrix} \otimes \boldsymbol{I}_M,$$

$$\boldsymbol{T}^{(\theta)} = \begin{bmatrix} 1 & 1 & \vdots & & & \boldsymbol{0}_{2\times 11} & & \\ 0 & 1 & \vdots & & & & & \\ \ldots & \ldots & \vdots & \ldots & \ldots & \ldots & \ldots & \ldots \\ & & \vdots & -1 & -1 & \ldots & -1 & -1 \\ & & \vdots & 1 & 0 & \ldots & 0 & 0 \\ \boldsymbol{0}_{11\times 2} & & \vdots & 0 & 1 & \ldots & 0 & 0 \\ & & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ & & \vdots & 0 & 0 & \ldots & 1 & 0 \end{bmatrix} \otimes \boldsymbol{I}_M.$$

The transformed survey error process $\{\boldsymbol{e}_t^*\}$ is assumed to follow an M-dimensional vector autoregressive moving average process (VARMA), defined by $\boldsymbol{\Phi}(B)\boldsymbol{e}_t^* = \boldsymbol{\Theta}(B)\boldsymbol{a}_t$, with mean vector $E(\boldsymbol{e}_t^*) = \boldsymbol{0}$ and

$$\boldsymbol{\Theta}(B) = \boldsymbol{I} - \boldsymbol{\Theta}_1 B - \ldots - \boldsymbol{\Theta}_q B^q,$$

$$\boldsymbol{\Phi}(B) = \boldsymbol{I} - \boldsymbol{\Phi}_1 B - \ldots - \boldsymbol{\Phi}_p B^p,$$

where $\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_p, \boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_q$ are coefficient matrices and $\boldsymbol{a}_t$ is an M-dimensional white noise random vector with zero mean and covariance structure:

$$E(\boldsymbol{a}_t \boldsymbol{a}'_{t-h}) = \begin{cases} \sum_a & h = 0, \\ \boldsymbol{0} & h \neq 0. \end{cases}$$

The cross-covariance matrix function for the VARMA process $\{\boldsymbol{e}_t^*\}$, (see Wei 1993, page 333), is given by:

$$\boldsymbol{\Gamma}_{e^*}(h) = \mathrm{COV}(\boldsymbol{e}_{t-h}^*, \boldsymbol{e}_t^*) = E(\boldsymbol{e}_{t-h}^* \boldsymbol{e}_t^{*\prime}),$$

where $\{\boldsymbol{\Gamma}_{e^*}(h)\}_{mj} = \gamma_{e^*,mj}(h) = \mathrm{COV}(e_{m,t-h}^*, e_{jt}^*)$, and the cross-correlation function for the vector process is defined as:

$$\boldsymbol{P}_{e^*}(h) = \boldsymbol{D}_{e^*}^{-1/2} \boldsymbol{\Gamma}_{e^*}(h) \boldsymbol{D}_{e^*}^{-1/2},$$

where

$$\boldsymbol{D}_{e^*} = \mathrm{diag}(\gamma_{e^*,11}(0), \ldots, \gamma_{e^*,MM}(0)).$$

The state space representation of VARMA models can be found in Reinsel (1993, section 7.2). The separate models for the transformed signal and sampling errors can be cast into a unique state space model, see Silva (1996, Chapter 8) for details.

4. ESTIMATION FROM THE TRANSFORMED DATA

As in previous sections, we distinguish between the estimation of the structure of the surveys errors, the noise, and the estimation of the covariances of the basic structural model. Once these are obtained, we employ the Kalman filter to get estimates of the trend and seasonals which determine the signal. Before carrying out the signal extraction, the VARMA model for the survey errors must be identified.

The model specification for the error process $\{\boldsymbol{e}_t^*\}$ depends on the sampling design, particularly on the level of sample overlap between occasions, and also on data availability. Many authors have considered the problem of modelling the sampling error process in a univariate framework, see, for example, Scott and Smith (1974), Pfeffermann (1989, 1991), Bell and Hillmer (1990), Binder and Dick (1989), Tiller (1989, 1992), Pfeffermann and Bleuer (1993), Binder, Bleuer and Dick (1993), Pfeffermann, Bell and Signorelli (1996) and Pfeffermann, Feder and Signorelli (1998). However, in all of these cases the authors are working with the original data instead of the transformed data. After transformation, it is difficult to carry out a full primary analysis based on individual observations, see Silva (1996, Chapter 7).

Many repeated surveys are based on a rotating panel design in which K panels of sampling units are investigated at each survey round (time point) and panels are replaced in a systematic manner, according to the rotating pattern of the survey design. In these surveys, elementary design unbiased estimates $y_t^{(k)}$, $k = 1, \cdots, K$, for the population parameter θ_t, can be obtained from each rotation group. A rotation group is a set of sampling units that joins and leaves the sample at the same time.

In a two-stage survey, in which the primary sampling units (enumeration areas) remain in the sample for all survey occasions, the replacement of panels of households (second-stage units) is ordinarily carried out within geographical regions defined by mutually exclusive groups of enumeration areas. Note that a survey with K panels produces K streams of estimates, where a stream is a time series of all sample estimates based on samples from the same enumeration area, that is, is a time series of elementary estimates.

Pfeffermann, Bell and Signorelli (1996) and Pfeffermann, Feder and Signorelli (1998) show how to estimate the autocorrelation of the sampling error process for univariate data, before transformation, using the so-called pseudo-errors, defined as:

$$\tilde{e}_t^{(k)} = y_t^{(k)} - y_t, \qquad (12)$$

where $y_t = 1/K \sum_{k=1}^{K} y_t^{(k)}$. If there is no rotation bias, it follows that:

$$\tilde{e}_t^{(k)} = e_t^{(k)} - e_t, \qquad (13)$$

thus contrasts in $y_t^{(k)}$ are contrasts in the panel sampling errors $e_t^{(k)}$.

For the compositional case we apply, for each elementary estimate, the transformation $v_t^{(k)} = a_m(y_t^{(k)}) = (v_{1t}^{(k)}, \ldots, v_{Mt}^{(k)})'$ which has as its m^{th} component, $(m = 1, \ldots, M)$:

$$v_{mt}^{(k)} = \log\left(\frac{y_{mt}^{(k)}}{y_{M+1,t}^{(k)}}\right) = \log\left(\frac{\theta_{mt}}{\theta_{M+1,t}}\right) + \log\left(\frac{u_{mt}^{(k)}}{u_{M+1,t}^{(k)}}\right). \qquad (14)$$

From (14), a vector model for the k^{th} series of transformed elementary estimates can be written as:

$$v_t^{(k)} = \theta_t^* + e_t^{*(k)}, \qquad (15)$$

with $e_t^{*(k)} = (e_{1t}^{*(k)}, \ldots, e_{Mt}^{*(k)})'$ and $e_{mt}^{*(k)} = \log(u_{mt}^{(k)} / u_{M+1,t}^{(k)})$, for $(m = 1, \ldots, M)$. Hence, from (15), M-dimensional time series of transformed pseudo-errors can be constructed from deviations of the transformed rotation group estimates about their overall mean. The transformed pseudo-errors for the k^{th} rotation group are defined as:

$$\tilde{e}_t^{*(k)} = (\tilde{e}_{1t}^{*(k)}, \ldots, \tilde{e}_{Mt}^{*(k)})' = v_t^{(k)} - v_t$$

$$= (v_{1t}^{(k)} - v_{1t}, \ldots, v_{Mt}^{(k)} - v_{Mt})', \qquad (16)$$

where $v_t = 1/K \sum_{k=1}^{K} v_t^{(k)}$. Note, in addition, that $\tilde{e}_t^{*(k)} = e_t^{*(k)} - e_t^*$.

From (14) and (15), it becomes clear that the framework introduced by Pfeffermann, Bell and Signorelli (1996) can also be applied to the transformed model.

The cross-correlation matrices of the transformed sampling errors can be obtained by averaging the cross-covariances matrices of the transformed pseudo-errors as follows (for details see Silva 1996, Chapter 7):

$$P_{\tilde{e}^*}(h) = \left[\sum_{k=1}^{K} D_{\tilde{e}^*}^{(k)}\right]^{-1/2} \left[\sum_{k=1}^{K} \Gamma_{\tilde{e}^*}^{(k)}(h)\right] \left[\sum_{k=1}^{K} D_{\tilde{e}^*}^{(k)}\right]^{-1/2}, \qquad (17)$$

where

$$\Gamma_{\tilde{e}^*}^{(k)}(h) = \mathrm{COV}(\tilde{e}_{t-h}^{*(k)}, \tilde{e}_t^{*(k)}) = E(\tilde{e}_{t-h}^{*(k)} \tilde{e}_t^{*(k)\prime}),$$

with

$$\{\Gamma_{\tilde{e}^*}^{(k)}(h)\}_{mj} = \mathrm{COV}(\tilde{e}_{m,t-h}^{*(k)}, \tilde{e}_{jt}^{*(k)}) = \gamma_{\tilde{e}^* mj}^{(k)}(h)$$

and

$$D_{\tilde{e}^*} = \mathrm{diag}(\gamma_{\tilde{e}^* 11}^{(k)}(0), \ldots, \gamma_{\tilde{e}^* MM}^{(k)}(0)).$$

Once the correlation matrices $P_{\tilde{e}^*}(h)$, $h = 1, 2, \ldots$ have been estimated, a VARMA model to represent the transformed survey error process can be selected and estimates of the respective parameter matrices can be computed, provided the series of transformed pseudo-errors are available. Then, as described in section 3, a state space model for representing the transformed signal and sampling errors can be defined and the Kalman filter equations can be used to get filtered and smoothed estimates for the unobservable components. The application of the Kalman Filter requires the estimation of the unknown hyperparameters (the covariance matrices $\Sigma_l, \Sigma_r, \Sigma_s, \Sigma_I, \Sigma_a$) and the estimation of the initial state vector and the respective covariance matrices.

Having addressed the issue of how to model the survey estimates in a compositional framework and how to identify the time series model for the transformed sampling errors, the following section presents the results of an empirical study using compositional data from the Brazilian Labour Force Survey.

5. MODELLING COMPOSITIONAL TIME SERIES IN THE BRAZILIAN LABOUR FORCE SURVEY

The Brazilian Labour Force Survey (BLFS) collects monthly information about employment, hours of work, education and wages together with some demographic information. It classifies the survey respondents, aged 15 and over, according to their employment status in the week prior to the interview into three main groups: employed, unemployed and not in the labour force, following the International Labour Organization (ILO) definitions. The survey targets the population living at the six major metropolitan areas in the country. The BLFS is a two-stage sample survey in which the primary sampling units (psu) are the census enumeration areas (EA) and the second-stage units (ssu) are the households. The primary sampling units

are selected with probabilities proportional to their sizes and then a fixed number of households is selected from each sampled EA by systematic sampling. All household members within the selected households are enumerated. The primary sampling units remain the same for a period of roughly 10 years (as in a master sample). New primary sampling units are selected when information from a new population census becomes available.

In addition, the BLFS is a rotating panel survey. For any given month the sample is composed of four rotation groups of mutually exclusive sets of primary sampling units. The rotation pattern applies to panels of second-stage units (households). Within each rotation group a panel of households stays in the sample for four successive months, is rotated out for the following 8 months and then is sampled again for another spell of four successive months. Each month one panel is rotated out of the sample. The substituting panel can be a new panel or one that has already been observed for the first four months period. Note that the 4-8-4 rotation pattern induces a complex correlation structure for the sampling errors over time and that there is a 75% overlap between two successive months.

The empirical work was carried out using data from the São Paulo metropolitan area covering the period from January 1989 to September 1993 (57 observations). The quantities of interest are the proportions of employed, unemployed and not in the labour force, and also the unemployment rate. Using the monthly individual observations, the series of sample estimates and their respective estimated standard errors were computed using data of each specific survey round and standard estimators. For each month, two sets of estimates were obtained. The direct sample estimates, derived from the complete data collected at a given month and four elementary estimates, each based on data from a single rotation group. The panel estimates are used to estimate the sampling error autocorrelations and to help to identify the time series model for the sampling errors.

In this study the observed composition has $M + 1 = 3$ components and the time series is defined as the sequence of vectors $y_t = (y_{1t}, y_{2t}, y_{3t})'$, where:

y_{1t} is the estimated proportion of unemployed persons in month t;

y_{2t} is the estimated proportion of employed persons in month t;

y_{3t} is the estimated proportion of persons not in the labour force in month t.

The model for the BLFS must incorporate the special features of the data. Firstly, it is a compositional time series belonging to the Simplex S^2 at each time t. Secondly, the time series are subject to sampling errors. Following the approach in section 2, we first map the composition onto $\mathbb{R}^2$ using the additive logratio transformation with y_{3t} as the reference category. As y_t is a vector of sample estimates, it can be modelled as in equation 1 and the vector model for the transformed series is given by equation 9. Then, the transformed composition is modelled using a multivariate state space model that accounts for the autocorrelations between the sampling errors. Finally, the model based estimates are transformed back to the original space. Figure 1 displays the series of transformed compositions.

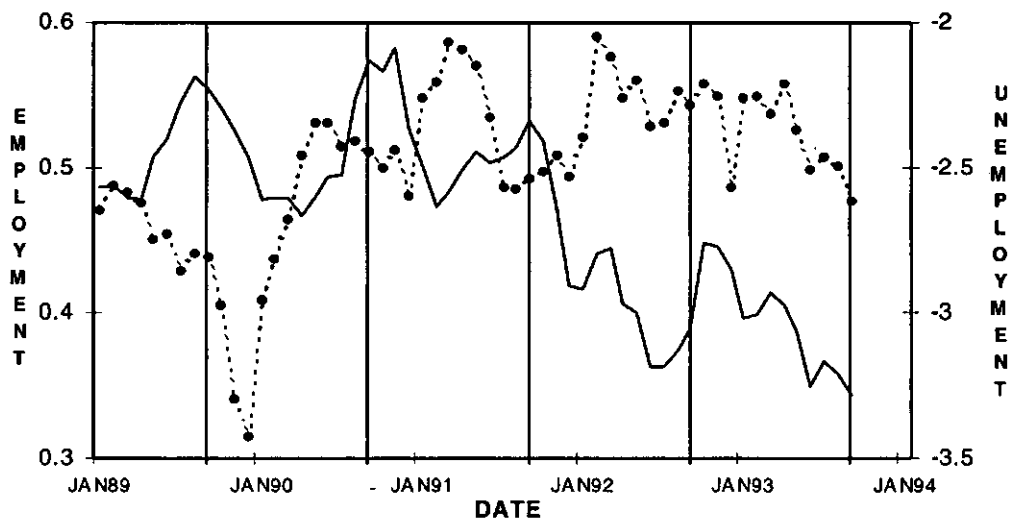

Figure 1. Brazilian Labour Force Series - SÃO PAULO Transformed Compositions

The model for the transformed sample estimates v_t is composed of a bivariate model for the transformed signal θ_t^*, describing how the transformed population quantities evolve in time, and a bivariate model representing the time series relationship between transformed sampling errors e_t^*. The transformed signal process $\{\theta_t^*\}$ is assumed to follow the bivariate basic structural model (equation 11) as described in section 3. As mentioned before, a VARMA model to represent the sampling error series was used. The correlation structure of the transformed sampling errors was estimated using the transformed pseudo-errors as in equation 16. In addition, estimates of the partial lag correlation matrices for $\{e_t^*\}$ were computed using a recursive algorithm provided in Wei (1993, pages 359-362). A program in SAS-IML which gives the corresponding schematic representations (Tiao and Box 1981) and a statistical test to help establish the order of the vector process was developed. The form of the correlation matrices and the results for the statistical test, available in Silva (1996), indicate that a VAR(1), a VAR(2) or a VARMA(1,1) model could be used to represent the transformed sampling error process. In the event, the VARMA(1,1) was chosen because it yields smaller standard errors for estimates of the unemployment rate. The parameter estimates for this model were obtained from the relationship between the cross-covariance function and the parameter matrices given in Wei (1993, pages 346-347). The VARMA(1,1) fitted for $\{e_t^*\}$ is given by:

$$\begin{bmatrix} e_{1t}^* \\ e_{2t}^* \end{bmatrix} = \begin{bmatrix} 0.7347 & 0.2414 \\ -0.9224 & -0.2072 \end{bmatrix} \begin{bmatrix} e_{1,t-1}^* \\ e_{2,t-1}^* \end{bmatrix} - \begin{bmatrix} 0.3162 & 0.2590 \\ -0.7666 & -0.2749 \end{bmatrix} \begin{bmatrix} a_{1,t-1} \\ a_{2,t-1} \end{bmatrix} + \begin{bmatrix} a_{1t} \\ a_{2t} \end{bmatrix},$$

with

$$\hat{\Sigma}_a = \begin{bmatrix} 0.0001723 & 0.0003476 \\ 0.0003476 & 0.0051660 \end{bmatrix} \qquad (18)$$

Having put the combined model for the transformed survey estimates into the state space form, the Kalman Filter equations can be used to get filtered and smoothed estimates for the unobservable components. Note that the estimation of the model for the transformed sampling errors (equation 18) was implemented outside the Kalman Filter. The application of the Kalman Filter requires the estimation of the unknown hyperparameters (the covariances), the initial state vector and respective covariance matrix. Assuming that the disturbances $\eta_t^{(\theta)}$, a_t and I_t are normally distributed, the log-likelihood function of the (transformed) observations can be expressed via the prediction error decomposition (for details see Harvey 1989). Estimates for the model covariances were obtained by maximum likelihood, applying a quasi-Newton optimization technique. A computer program to implement the maximization procedure was developed using the optimization routine NLPQN from SAS-IML.

The initialization of the Kalman filter was carried out using a combination of a diffuse and proper priors. Following this approach, the non-stationary components $(\alpha^{(\theta)})'$ of the state vector were initialized with very large error variances and the respective components of the initial state vector were taken as zero. The stationary components $(e_{1t}^*, e_{2t}^*)'$ were initialized by the corresponding unconditional mean and variance.

When fitting the model, the estimated covariance matrices obtained for the slope and seasonal components were very small and could be set to zero. This implies that the seasonals are assumed to be deterministic and that the slope is assumed to be fixed, giving rise to a local level model with a drift and non-stochastic seasonals for the signal. Indeed, as pointed out by Koopman, Harvey, Doornik and Shephard (1995, page 39), when the number of years considered in the analysis is small, it seems reasonable to fix the seasonals since there is not enough data to allow the estimation of a changing pattern. The fact that a fixed seasonal pattern is validated by the estimation process is a satisfactory feature of the modelling procedure. In addition, the estimated covariance matrix of the irregular component was also found to be very small (and hence undetectable) in comparison to the sampling error and so, as expected, in the presence of relatively large sampling errors, there was no need to include irregular components in the model for the transformed signal. The parameter estimates and respective asymptotic errors (displayed in parenthesis) are presented in Table 1.

Table 1
Estimates for the Hyperparameters and Standard Errors

Model	$\hat{\Sigma}_l \times 10^{-4}$ (2)	$\hat{\Sigma}_r = \hat{\Sigma}_s = \hat{\Sigma}_I$
BSM + VARMA (1,1) (1)	$\begin{bmatrix} 2.78 & \mathbf{0.12} \\ (0.91) & \\ 1.95 & 87.0 \\ (3.55) & (27.10) \end{bmatrix}$	$\begin{bmatrix} 0 & - \\ 0 & 0 \end{bmatrix}$

(1) Local level model with drift and fixed seasonals for the signal.
(2) Upper-triangular contains correlation.

To evaluate the model performance, empirical distributions of the standardized residuals were compared with a standard normal distribution to verify the assumption that the innovations $(v_t - \hat{v}_{t|t-1})$ are normal deviates. Examination of corresponding normal plots revealed no departure from normality. In addition, we also computed the auto-

correlations of the innovations, which were close to zero, further validating the model.

Predictions for y_{mt} and estimates for θ_{mt} are computed by applying the additive logistic transformation (equation 5) to predictions $\hat{\nu}_{t|t-1}$ and smoothed estimates $\hat{\theta}^*_{t|T}$ for the transformed series and signal, respectively. This transformation maps these estimates onto S^2, guaranteeing that they satisfy the boundedness constraints.

Unfortunately, although $\hat{L}^*_{t|T}$ and $\hat{S}^*_{t|T}$ can be obtained from $\hat{\theta}^*_{t|T}$, it is not straightforward to obtain estimates for the structural unobservable components of the original signal θ_t, such as $\hat{L}_{t|T}$ and $\hat{S}_{t|T}$, However, if a multiplicative model with no irregular component is assumed for $\{\theta_{mt}\}$, such that:

$$\theta_{1t} = L_{1t} S_{1t}\,,\ \theta_{2t} = L_{2t} S_{2t}\,,\ \theta_{3t} = L_{3t} S_{3t}, \quad (19)$$

where L_{mt} and S_{mt}, for $m = 1,2,3$ represent the trend and seasonal components of the unobservable signals, then applying an additive logratio transformation to θ_t results in:

$$\log(\theta_{mt}/\theta_{3t}) = \log\left(\frac{L_{mt}\,S_{mt}}{L_{3t}\,S_{3t}}\right)$$

$$= \log\left(\frac{L_{mt}}{L_{3t}}\right) + \log\left(\frac{S_{mt}}{S_{3t}}\right), m = 1,2. \quad (20)$$

This can be rewritten as:

$$\theta^*_{mt} = L^*_{mt} + S^*_{mt}, \quad (21)$$

with $L^*_{mt} = \log(L_{mt}/L_{3t})$ and $S^*_{mt} = \log(S_{mt}/S_{3t})$. Thus, the use of a basic structural model for $\{\theta^*_t\}$ corresponds to the case in which the underlying model for $\{\theta_t\}$ decomposes the original signal into its trend and seasonal components in a multiplicative way. For deriving estimates, either filtered or smoothed, for L_{mt} note that:

$$\exp(L^*_{1t}) = L_{1t}/L_{3t},\ \exp(L^*_{2t}) = L_{2t}/L_{3t}. \quad (22)$$

To recover L_{1t}, L_{2t}, L_{3t}, in (22), it is necessary to assume an explicit relationship between these unobservable components based on model (19). By doing this, a third equation can be added to the system in (22) and an estimate of the original series components can be obtained. Note that the system has three unknowns for just two equations. In this case, it is quite natural to assume that the level components sum to one across the series, being also bounded between zero and one. Hence, trend estimates for the original series can be obtained solving :

$$\begin{cases} \exp(L^*_{1t}) = L_{1t}/L_{3t}, \\ \exp(L^*_{2t}) = L_{2t}/L_{3t}, \\ L_{1t}+L_{2t}+L_{3t} = 1, \end{cases} \quad (23a)$$

which results in

$$L_{mt} = \frac{\exp(L^*_{mt})}{1 + \sum_{k=1}^{2} \exp(L^*_{kt})},\ m = 1,2;$$
$$L_{3t} = \frac{1}{1 + \sum_{k=1}^{2} \exp(L^*_{kt})}. \quad (23b)$$

As there is no irregular component in model (19) the seasonally adjusted figures are given by the trend estimates in (23). Therefore, the smoothed estimates for the trend of the original series of proportions are obtained by applying the additive logistic transformation to $\hat{L}^*_{t|T}$. Consequently, estimates for the seasonal components of the original proportions can be computed as :

$$\hat{S}_{m,t|T} = \hat{\theta}_{m,t|T}/\hat{L}_{m,t|T},\ m = 1,2,3.$$

For labour force surveys, an important issue is the estimation of the unemployment rate series (as opposed to unemployment proportions) and also the production of the corresponding seasonally adjusted figures. Recall that θ_{1t} and θ_{2t} represent the unknown population proportions of unemployed and employed people, respectively. Using these proportions, the unknown unemployment rate at time is t defined as

$$R_t = \frac{\theta_{1t}}{\theta_{1t}+\theta_{2t}} = \frac{1}{\left(1 + \frac{\theta_{2t}}{\theta_{1t}}\right)} = \left(\frac{\theta_{2t}}{\theta_{1t}} + 1\right)^{-1}. \quad (24)$$

Based on model (11), trend estimates for the unemployment rate can be obtained by simply replacing θ_{mt} by L_{mt}, $m = 1,2$, in equation 24. In conclusion, the methodology developed in this section provides signal (and trend) estimates that are bounded between zero and one and satisfy the unit-sum constraint. It also provides estimates for the seasonal and trend components of series comprising ratios of the original proportions which is a useful feature.

Figure 2 presents the design-based estimates and the model-dependent estimates for the proportion of unemployed persons, for the time period January 1989 to September 1993. The model-dependent estimates are the smoothed estimates which use all the data for the whole sample period. As can be seen from the graph, the signal estimates behave similarly to the design-based estimates although some of the sharp turning points in the series have been smoothed out.

Model-dependent trend estimates were obtained by fitting the basic structural model defined for the signal process when sampling error variation was modelled as a VARMA(1,1). These estimates were compared with the estimates produced by the familiar X-11 procedure. Figure 3 displays the trend produced for the unemployment rate

series by both methods together with the estimates obtained by fitting a standard basic structural model which does not account for sampling error variation.

The trend produced by our model is smoother, suggesting that the model-dependent procedure succeeds in removing the fluctuations induced by the sampling errors.

In addition, model-dependent estimates for the seasonal effects of the original compositions were also obtained from the multivariate modelling procedure which accounts for two very important features of the data, namely the compositional constraints and the presence of sampling errors.

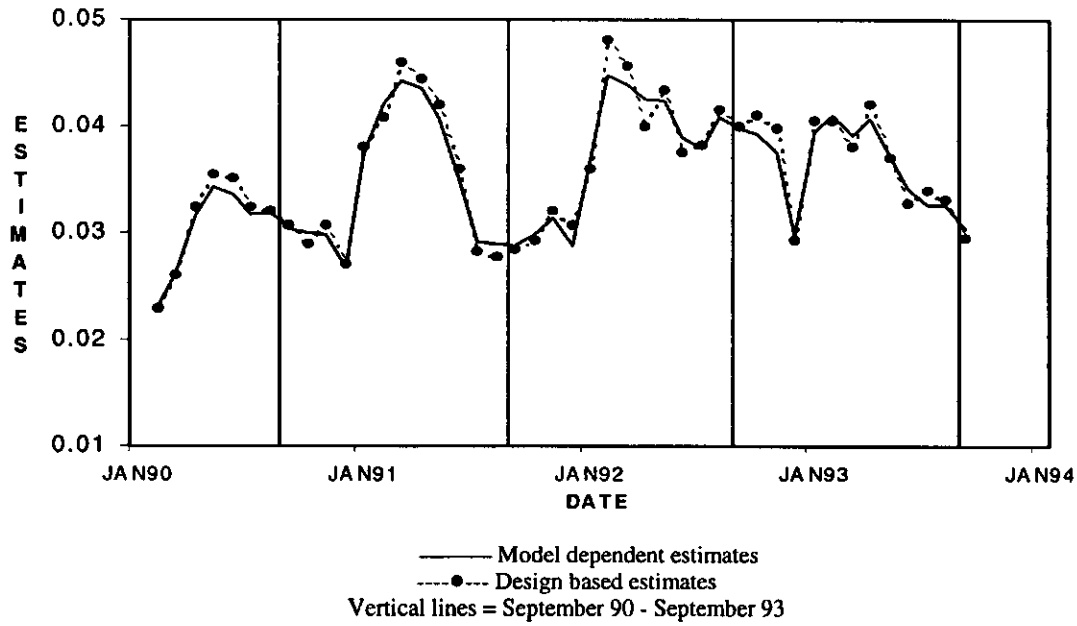

Figure 2. Brazilian Labour Force Series - SÃO PAULO Design Based and Model Dependent Estimates Proportion of Unemployed Persons

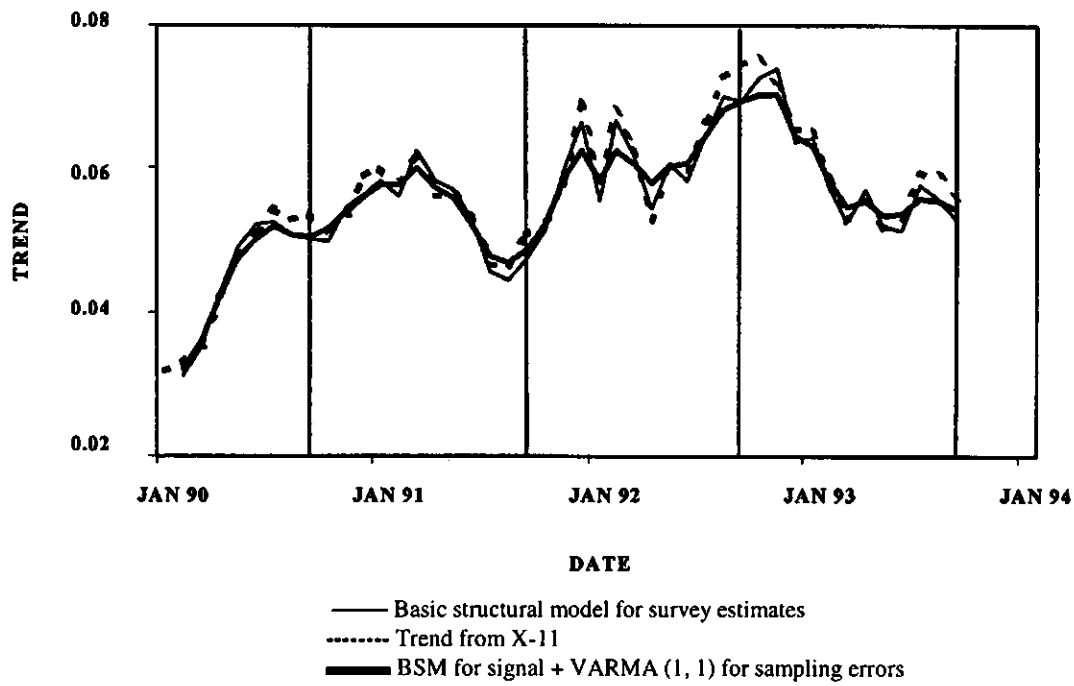

Figure 3. Brazilian Labour Force Series - SÃO PAULO Trend Estimates for the Unemployed Rates Series

6. CONCLUSIONS

This paper proposes a state space approach for modelling compositional time series from repeated surveys. The important feature of the proposed methodology is that it provides bounded predictions and signal estimates of the parameters in a composition, satisfying the unity-sum constraint, while taking into account the sampling errors. This is accomplished by mapping the compositions from the Simplex onto Real space using the additive logratio transformation, modelling the transformed data employing multivariate state space models, and then applying the additive logistic transformation to obtain estimates in the original scale.

The empirical work using data from the Brazilian Labour Force Survey demonstrates the usefulness of this modelling procedure in a genuine survey situation, showing that it is possible to model the multivariate system and obtain estimates for all the relevant components. The results of the empirical work also show that smoother trends and fixed seasonals are obtained from a model which explicitly accounts for the sampling errors, when compared with estimates produced by X-11. In addition, because the model-dependent estimators combine past and current survey data, the standard deviations of these estimates are in general lower that the standard deviations of the design-based estimators, as shown in Silva (1996, Chapter 8).

One drawback of the proposed procedure is that although confidence regions for the original compositional vector can be constructed based on the model-dependent estimates by using the additive logistic normal distribution, confidence intervals for the individual proportions are not readily available. Such intervals could be obtained from marginal distributions of the additive logistic normal distribution, but these can only be evaluated by integrating out some of the elements of the compositional vector and, as pointed out by Brunsdon (1987, page 135), this produces intractable expressions.

Under a state space formulation a wide variety of models is available to represent the multivariate signal and noise processes, which is a great benefit of this modelling procedure. The application of the method to different data sets is recommended. Further empirical research should also consider situations where the composition lies on a Simplex with dimensions higher than two and/or with compositions evolving close to the boundaries of the interval [0.1]. In addition, a better insight into the performance of the modelling procedure may be gained by applying the method to simulated data, for which the "true" underlying models are known. The models considered here can also be extended to incorporate rotation group bias effects and explanatory variables.

ACKNOWLEDGEMENTS

This research was supported by CAPES-Brazil and IBGE-Brazil and by a grant from the Economic and Social Research Council of the UK under its Analysis of Large and Complex Datasets Programme. The authors wish to thank the referees and Prof. Danny Pfeffermann for the suggestions that led to many improvements in the paper. Thanks are also due to Dr. Harold Mantel for his encouragement towards the preparation of the final version.

REFERENCES

AITCHISON, J. (1986). *The Statistical Analysis of Compositional Data*. New York: Chapman and Hall.

AITCHISON, J., and SHEN, S.M. (1980). Logistic-Normal distributions: some properties and uses. *Biometrika*, 67, 261-272.

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

BINDER, D.A., and HIDIROGLOU, M.A. (1988). Sampling in time. In *Handbook of Statistics*, (Eds., P.R. Krishnaiah and C.R. Rao). Elsevier Science, 6, 187-211.

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

BINDER, D.A., BLEUER, S.R. and DICK, J.P. (1993). Time series methods applied to survey data. *Proceedings of the 49° International Statistical Institute Session*, 1, 327-344.

BLIGHT, B.J.N., and SCOTT, A.J. (1973). A stochastic model for repeated surveys. *Journal of the Royal Statistical Society, B*, 35, 61-68.

BRUNSDON, T.M.(1987). Time Series Analysis of Compositional Data. Unpublished Ph.D. Thesis. University of Southampton.

BRUNSDON, T.M., and SMITH, T.M.F. (1998). The time series analysis of compositional data. *Journal of Official Statistics*, 14, 3, 237-253.

DE JONG, P.(1988). The likelihood for a state space model. *Biometrika*, 75, 165-169.

DE JONG, P. (1989). Smoothing and interpolation with the state space model. *Journal of the American Statistical Society*, 84, 1085-1088.

DE JONG, P. (1991). The diffuse Kalman filter. *The Annals of Statistics*, 19, 1073-1083.

FERNANDEZ, F.J.M., and HARVEY, A.C. (1990). Seemingly unrelated time series equations and a test for homogeneity. *Journal of Business and Economic Statistics*, 8, 1, 71-81.

GURNEY, M., and DALY, J.F. (1965). A multivariate approach to estimation in periodic sample surveys. *Proceedings of the American Statistical Association, Social Statistics Section*, 242-257.

HARRISON, P.J., and STEVENS, C.F. (1976). Bayesian forecasting. *Journal of the Royal Statistical Society, B*, 38, 205-47.

HARVEY, A.C. (1986). Analysis and generalisation of a multivariate exponential smoothing model. *Management Science*, 32, 374-380.

HARVEY, A.C. (1989). Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge University Press. Cambridge.

HARVEY, A.C. (1993). *Time Series Models*. Second Edition. Harvester Wheatsheaf. London.

HARVEY, A.C., and PETERS, S. (1984). Estimation Procedures for Structural Time Series Models. London School of Economics. Mimeo.

HARVEY, A.C., and SHEPHARD, N. (1993). Structural time series models. In *Handbook of Statistics*, (Eds. S. Maddala, C.R. Rao and H.D. Vinod). Elsevier Science Publishers, 11, 261-302.

HARVEY, A.C., and CHUNG, C.(2000). Estimating the underlying change in unemployment in the UK. *Journal of the Royal Statistical Society, A*, 163, Part 3, 303-339.

IBGE (1980). Metodologia da Pesquisa Mensal de Emprego 1980. Relatórios Metodológicos. Fundação Instituto Brasileiro de Geografia e Estatística. Rio de Janeiro.

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

KOOPMAN, S.J., HARVEY, A.C., DOORNIK, J.A. and SHEPHARD, N. (1995). *STAMP 5.0 - Structural Time Series Analyser, Modeller and Predictor*. Chapman & Hall. London.

MITTNIK, S. (1991). Derivation of the unconditional state covariance matrix for exact-likelihood estimation of ARMA models. *Journal of Economic Dynamics and Control*, 15, 731-740.

PFEFFERMANN, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business and Economic Statistics*, 9, 163-177.

PFEFFERMANN, D., BURCK, L. and BEN-TUVIA, S. (1989). A time series model for estimating housing price indexes adjusted for changes in quality. *Proceedings of the International Symposium on Analysis of Data in Time*, 43-55.

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

PFEFFERMANN, D., and BLEUER, S.R. (1993). Robust joint modelling of labour force series of small areas. *Survey Methodology*, 19, 149-164.

PFEFFERMANN, D., BELL, P. and SIGNORELLI, D. (1996). Labour force trend estimation in small areas. *Proceedings of the Annual Research Conference, Bureau of the Census*, 407- 431.

PFEFFERMANN, D., FEDER, M. and SIGNORELLI, D. (1998). Estimation of autocorrelations of survey errors with applications to trend estimation in small samples. *Journal of Business and Economics Statistics*, 16, 339-348.

QUINTANA, J.M., and WEST, M. (1988). Time series analysis of compositional data. *Journal of Bayesian Statistics*, (Eds. J.H. Bernado, M.A. Degroot and A.F.M. Smith). Oxford University Press, 3.

REINSEL, G.C. (1993). *Elements of Multivariate Time Series Analysis*. Springer-Verlag.

SAS INSTITUTE INC. (1995). *SAS/IML Software: Changes and Enhancements through Release 6.11*. SAS institute Inc. Cary, NC.

SCOTT, A.J., and SMITH, T.M.F. (1974). Analysis of repeated surveys using time series methods. *Journal of the American Statistical Association*, 69, 674-678.

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

SHEPHARD, N.G., and HARVEY, A.C. (1989). Tracking the Level of Support for the Parties During General Election Campaigns. Mimeo. Dept. of Statistics, London School of Economics.

SILVA, D.B.N. (1996). Modelling Compositional Time Series From Repeated Surveys. Unpublished PhD Thesis. University of Southampton. UK.

SINGH, A.C., and ROBERTS, G.R. (1992). State space modelling of cross-classified time series of counts. *International Statistical Review*, 60, 321-335.

SMITH, T.M.F., and BRUNSDON, T.M. (1986). Time Series Methods for Small Areas. Unpublished Report. University of Southampton.

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

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

TILLER, R.B.(1992). Time series modelling of sample data from the U.S. Current Population Survey. *Journal of Official Statistics*, 8, 2, 149-166.

WALLIS, F. (1987). Time series analysis of bounded economic variables. *Journal of Time Series Analysis*, 8, 115-23.

WEI, W.W.S. (1993). *Time Series Analysis - univariate and multivariate methods*. Addison-Wesley.

WEST, M., and HARRISON, J. (1989). *Bayesian Forecasting and Dynamic Models*. Springer-Verlag.

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following persons who have served as referees during 2001. An asterisk indicates that the person served more than once.

M. Axelson, *Örebro University, Sweden*
M. Bankier, *Statistics Canada*
K. Brewer, *Australian National University*
Moon Jung Cho, *U.S. Bureau of Labor Statistics*
R. Chambers, *University of Southampton*
* S. Chowdhury, *Westat, Inc.*
M. P. Cohen, *U.S. Bureau of Transportation Statistics*
G. Datta, *University of Georgia*
P. Duchesne, *École des Hautes Études Commerciales de Montréal*
F. Dupont, *INSEE*
M.R. Elliott, *University of Pennsylvania*
J.L. Eltinge, *U.S. Bureau of Labor Statistics*
* V. M. Estevao, *Statistics Canada*
M. Ghosh, *University of Florida*
B. Graubard, *National Cancer Institute*
R. Harter, *National Opinion Research Center*
M.A. Hidiroglou, *Statistics Canada*
B. Hulliger, *Swiss Federal Statistical Office*
D. Jang, *Mathematica Policy Research*
D. Kostanich, *U.S. Bureau of the Census*
P. Kott, *NASS*
* M. Kovačević, *Statistics Canada*
N. Laniel, *Statistics Canada*
M.D. Larsen, *University of Chicago*
M. Latouche, *Statistics Canada*
* P. Lavallée, *Statistics Canada*
H. Lee, *Westat, Inc.*
J. Lent, *U.S. Bureau of Transportation Statistics*
* S. Lohr, *Arizona State University*
W. Lu, *Simon Fraser University*
J. Moloney, *Statistics Canada*
G. Montanari, *University of Perugia*
C. Perry, *NASS*
T.E. Raghunathan, *University of Michigan*
E. Rancourt, *Statistics Canada*
L.-P. Rivest, *Université Laval*
N. Schenker, *National Center for Health Statistics*
J. Sedransk, *Case Western University*
* A.C. Singh, *Research Triangle Institute*
K.P. Srinath, *ABT Associates*
B. Sutrahar, *Memorial Univesity*
* A. Théberge, *Statistics Canada*
R. Thomas, *Carleton University*
S. K. Thompson, *Pennsylvania State University*
C. Tucker, *U.S. Bureau of Labor Statistics*
* R. Valliant, *Westat, Inc.*
J. Waksberg, *Westat, Inc.*
C. Wu, *University of Waterloo*
Y. You, *Statistics Canada*
* W. Yung, *Statistics Canada*
* E. Zanutto, *University of Pennsylvania*

Acknowledgements are also due to those who assisted during the production of the 2001 issues: H. Laplante (Dissemination Division) and L. Perreault (Official Languages and Translation Division). Finally we wish to acknowledge C. Ethier, C. Larabie, and D. Lemire of Household Survey Methods Division, for their support with coordination, typing and copy editing.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 17, No. 2, 2001

Volume 17, No. 3, 2001

Volume 17, No. 4, 2001

All inquires about submissions and subscriptions should be directed to the Chief Editor:
Lars Lyberg, R&D Department, Statistics Sweden, Box 24 300, S - 104 51 Stockholm, Sweden.

CONTENTS TABLE DES MATIÈRES

Volume 29, No. 2, June/juin 2001

Volume 29, No. 3, September/septembre 2001

Volume 29, No. 4, December/décembre 2001

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue of *Survey Methodology* (Vol. 19, No. 1 and onward) as a guide and note particularly the points below. Accepted articles must be submitted in machine-readable form, preferably in WordPerfect. Other word processors are acceptable, but these also require paper copies for formulas and figures.

1. Layout

1.1 Manuscripts should be typed on white bond paper of standard size (8½ × 11 inch), one side only, entirely double spaced with margins of at least 1½ inches on all sides.
1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
1.4 Acknowledgements should appear at the end of the text.
1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

3.1 Avoid footnotes, abbreviations, and acronyms.
3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as "exp(·)" and "log(·)", etc.
3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
3.4 Write fractions in the text using a solidus.
3.5 Distinguish between ambiguous characters, (e.g., w, ω; o, O, 0; l, 1).
3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.