Catalogue no. 12-001-XWE
ISSN: 1492-0921

Catalogue no. 12-001-XPB
ISSN: 0714-0045

Survey Methodology

June 2013

Statistics Canada
Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at infostats@statcan.gc.ca,

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-877-287-4369

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

To access this product

This product, Catalogue no. 12-001-X, is available free in electronic format. To obtain a single issue, visit our website, www.statcan.gc.ca, and browse by “Key resource” > “Publications.”

This product is also available as a standard printed publication at a price of CAN$30.00 per issue and CAN$58.00 for a one-year subscription.

The following additional shipping charges apply for delivery outside Canada:

	Single issue	Annual subscription
United States	CAN$6.00	CAN$12.00
Other countries	CAN$10.00	CAN$20.00

All prices exclude sales taxes.

The printed version of this publication can be ordered as follows:

- Telephone (Canada and United States) 1-800-267-6677
- Fax (Canada and United States) 1-877-287-4369
- E-mail infostats@statcan.gc.ca
- Mail Statistics Canada
Finance
R.H. Coats Bldg., 6th Floor
150 Tunney's Pasture Driveway
Ottawa, Ontario K1A 0T6
- In person from authorized agents and bookstores.

When notifying us of a change in your address, please provide both old and new addresses.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under “About us” > “The agency” > “Providing services to Canadians.”

Published by authority of the Minister responsible for Statistics Canada

Cette publication est aussi disponible en français.

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard symbols

The following symbols are used in Statistics Canada publications:

.	not available for any reference period
..	not available for a specific reference period
...	not applicable
0	true zero or a value rounded to zero
0^s	value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p}	preliminary
^{r}	revised
x	suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E}	use with caution
F	too unreliable to be published
*	significantly different from reference category ($p < 0.05$)

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48 - 1984.
∞

Le papier utilisé dans la présente publication répond aux exigences minimales de l'"American National Standard for Information Sciences" – "Permanence of Paper for Printed Library Materials", ANSI Z39.48 - 1984.
∞

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is indexed in The ISI Web of knowledge (Web of science), 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. It is also covered by SCOPUS in the Elsevier Bibliographic Databases.

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 in electronic format. Authors are invited to submit their articles in English or French in electronic form, preferably in Word to the Editor, (smj@statcan.gc.ca, Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6). For formatting instructions, please see the guidelines provided in the journal and on the web site (www.statcan.gc.ca/SurveyMethodology).

Survey Methodology

A Journal Published by Statistics Canada

Volume 39, Number 1, June 2013

Contents

Regular Papers

Short Notes


ELECTRONIC PUBLICATIONS AVAILABLE AT
PUBLICATIONS ÉLECTRONIQUES DISPONIBLE À
www.statcan.gc.ca

Survey Methodology, June 2013
Vol. 39, No. 1, pp. 1-27
Statistics Canada, Catalogue No. 12-001-X

Objective stepwise Bayes weights in survey sampling

Jeremy Strief and Glen Meeden[1]

Abstract

Although weights are widely used in survey sampling their ultimate justification from the design perspective is often problematical. Here we will argue for a stepwise Bayes justification for weights that does not depend explicitly on the sampling design. This approach will make use of the standard kind of information present in auxiliary variables however it will not assume a model relating the auxiliary variables to the characteristic of interest. The resulting weight for a unit in the sample can be given the usual interpretation as the number of units in the population which it represents.

Key Words: Sample survey; Weights; Bayesian inference.

1 Introduction

Weights play an important role in the design based approach to survey sampling. In theory the weight assigned to an observed unit in a sample is the reciprocal of its selection probability and is interpreted as the number of units in the population which it represents. In practice, after a sample has been observed, the weights are often adjusted to make the sample better represent the population. These adjustments can be made to take into account population information not included in the design and for observations missing from the sample. Although such modifications of the design based weights are undoubtedly useful in some cases their ultimate theoretical justification is not so clear. Part of the confusion, we believe, comes from arguing unconditionally before the sample is taken, *e.g.*, the Horvitz-Thompson estimator is unbiased averaged over all possible samples, and then conditionally after the sample is in hand, by adjusting the designed based weights of the observed units in the sample. In particular, an overemphasis on the sampling design at the second or conditional stage can needlessly complicated matters. After the sample has been observed, we believe a better approach is to formally ignore the sampling design but use all the available information, including that

1. Jeremy Strief, Principal Statistician, Medtronic Energy and Component Center, Brooklyn Center, MN 55430. E-mail : jstrief@gmail.com; Glen Meeden, School of Statistics, University of Minnesota, Minneapolis, MN 55455. E-mail : glen@stat.umn.edu.

embedded in the design, to find a sensible set of weights. In this way of thinking a weight assigned to a unit can still be interpreted as the number of units in the population that it represents but it is no longer derived as an adjustment of its selection probability. How can this be done?

In the Bayesian approach information about the population is incorporated into a prior distribution. In theory, the prior can then be used to purposely select an optimal sample; however this is almost never done. After the sample is observed inferences are based on the posterior distribution of the unobserved units in the population given the values of the observed units in the sample. In most situations the posterior does not depend on how the sample was selected and hence the design plays no role at the inference stage. Bayes methods have been little used in practice because it is difficult to find prior distributions which reflect the common kinds of available prior information.

Many of the standard estimators can be given a stepwise Bayesian interpretation (Ghosh and Meeden 1997). In this approach, given any sample, inference is still based on a posterior distribution but the collection (for all possible samples) of the posteriors does not arise from a single prior but from a whole family of prior distributions. In the situation where one believes that the observed units are roughly exchangeable with the unobserved units the appropriate stepwise Bayes posterior distribution is the Polya posterior.

When prior information about population means and quantiles of auxiliary variables is available Lazar, Meeden and Nelson (2008) argued that the constrained Polya posterior, a generalization of the Polya posterior, is a sensible way to incorporate such prior information. Here we will show how the constrained Polya posterior can be used to define weights for the units in the sample. Although the resulting weights depend on the auxiliary variables they do not make explicit use of the sampling design.

In Section 2 we review the Polya posterior and in Section 3, the constrained Polya posterior. The two main ideas of the paper are given in the next two sections. In Section 4 we show how the constrained Polya posterior can be used to attached a weight to each unit in the sample and in

such a way that these weights do not depend directly on the sampling design. In Section 5 we introduce the weighted Dirichlet posterior as a companion to the constrained Polya posterior. It allows one to use the weights defined by the constrained Polya posterior to make inferences about population parameters through straight forward simulation. In Section 6 we compare the constrained Polya posterior weights to those used in the Horvitz-Thompson estimator. In Section 7 we consider several examples to see how the resulting weights preform in practice and show how the weighted Dirichlet posterior can be use to get an estimate of variance for an estimator without extensive computing. Section 8 contains some concluding remarks.

At first reading it will seem to some that the methods proposed here are very Bayesian because all of our inferences are based on "posterior" distributions. But as mentioned above, technically, our "posterior" distributions are not Bayes but stepwise Bayes. This means that operationally one can think of our posterior as being constructed after the sample has been observed. These constructed "posteriors" do not depend on subjective prior information or the sampling design but just use the observed sample values and objective and public information about the auxiliary variables. As we shall see this allows one to construct estimators of population parameters which are approximately unbiased under a variety of designs and have good frequentist properties. There are two important limitations of our work however. The first is that it only is applicable to single stage designs and the second is that it cannot correct for selection bias.

2 The Polya posterior

Let s be the set of labels of a sample of size n from a population of size N. For convenience we assume the members of s are $1, 2, \ldots, n$ and we also suppose that n / N is very small. Let $y = (y_1, y_2, \ldots, y_N)$ be the characteristic of interest and y_s be the observed sample values.

The Polya posterior is based upon Polya sampling from an urn. Polya sampling works as follows: suppose that the values from n observed or seen units are marked on n balls and placed

in urn 1. The remaining unseen $N - n$ units of the population are represented by $N - n$ unmarked balls placed in urn 2. One ball from each urn is drawn with equal probability, and the ball from urn 2 is assigned the value of the ball from urn 1. Both balls are then returned to urn 1. Thus at the second stage of Polya sampling, urn 1 has $n + 1$ balls and urn 2 has $N - n - 1$ balls. This procedure is repeated until urn 2 is empty, at which point the N balls in urn 1 constitute one complete simulated copy of the population. Any finite population quantity – means, totals, quantiles, regression coefficients – may now be calculated from the complete copy. For the population quantity of interest we may simulate K such complete copies and in each case calculate its value. The mean of these simulated values is the point estimate and an approximate 95% Bayesian credible interval is given by the 2.5% and 97.5% quantiles of the values.

One can check that under the Polya posterior the posterior expectation of the population mean is just the sample mean and the posterior variance is just $(n - 1) / (n + 1)$ times the usual design based variance of the sample mean under simple random sampling without replacement. The Polya posterior has a decision theoretic justification based on its stepwise Bayes nature. Using this fact many standard estimators can be shown to be admissible. Details can be found in Ghosh and Meeden (1997). The Polya posterior is the Bayesian bootstrap of Rubin (1981) applied to finite population sampling. Lo (1988) also discusses the Bayesian bootstrap in finite population sampling. Some early related work can be found in Hartley and Rao (1968) and Binder (1982).

For the sample unit i let p_i denote the proportion of units in a full, simulated copy of the population which have the value y_i. Ghosh and Meeden (1997) showed that under the Polya posterior $E(p_i) = 1 / n$. If we let

$$w_i = NE(p_i) = N / n$$

then w_i can be interpreted as the weight attached to unit i since it equals the average number of units in the population represented by unit i, under the Polya posterior. Recall that under simple random sampling without replacement n / N is the inclusion probability for each unit. Hence in

this case the usual frequentist weight, which is the reciprocal of the inclusion probability, and Polya posterior weight defined above agree.

So in situations of limited prior information the Polya posterior yields weights identical to frequentist weights derived from the design of simple random sampling without replacement. The Polya posterior justification for these weights does not depend explicitly on the design and would be appropriate anytime the sampler believes the observed and unobserved units in the population are roughly exchangeable.

We next address the issue of the relationship of the Polya posterior with usual bootstrap methods in finite population sampling. Both approaches are based on an assumption of exchangeability. Gross (1980) introduced the basic idea for the bootstrap. Assume simple random sampling without replacement and suppose it is the case that $N / n = m$ is an integer. Given a sample we create a good guess for the population by combining m replicates of the sample. By taking repeated random samples of size n from this created population we can study the behavior of an estimator of interest. Booth, Bulter and Hall (1994) studied the asymptotic properties of such estimators. Hu, Zhang, Cohen and Salvucci (1997) is an example where the sample was used to construct an artificial population and then repeated samples were drawn from the constructed population to construct an estimate of the variance of their estimator and to construct confidence intervals.

Note this is in contrast to the Polya posterior which considers the sample fixed and repeatedly generates complete versions of the population.

3 The constrained Polya posterior

We begin by recalling a well known approximation to the Polya posterior. If n / N is small then under the Polya posterior, $p = (p_1, \ldots, p_n)$ has approximately a Dirichlet distribution with a parameter vector of all ones, *i.e.*, it is uniform on the $n - 1$ dimensional simplex, where

$\sum_{j=1}^{n} p_j = 1$. It is usually more efficient to generate complete copies of the population using this approximation than the urn model described in the previous section. In addition this approximation will be useful when we consider the constrained Polya posterior, a generalization of the Polya posterior which arises when prior information about auxiliary variables are available to the sampler.

In many problems, in addition to the variable of interest, y, the sampler has in hand auxiliary variables for which prior information is available. A very common case is when the population mean of an auxiliary variable is known. More generally, we will assume that prior information about the population can be expressed by a set of linear equality and inequality constraints on a collection of auxiliary variables.

We assume that in addition to the characteristic of interest y there is a set of auxiliary variables $x^1, x^2, \ldots, x^m$. For unit i let

$$(y_i, x_i) = (y_i, x_i^1, x_i^2, \ldots, x_i^m)$$

be the vector of values for y and the auxiliary variables. We suppose that for any unit in the sample this vector of values is observed. We assume the prior information about the population can be expressed through a set of linear equality and inequality constraints on the population values of the auxiliary variables. For the set of possible values for a given auxiliary variable the coefficients defining a constraint will correspond to the proportions of units in the population taking on these values. We now illustrate this more precisely by explaining how we translate this prior information about the population to the observed sample values. Given a sample this will allow us to construct simulated copies of the population consistent with the prior information.

Given a sample s, for $i = 1, 2, \ldots, n$, let (y_i, x_i) be the observed values which, for simplicity, we assume are distinct. Let p_i be the proportion of units which are assigned the value (y_i, x_i) in a simulated complete copy of the population. Any linear constraint on the population value of an auxiliary variable translates in an obvious way to a linear constraint on these observed values. For example, if the population mean of x^1 is known to be less than or equal to some

value, say b_1, then for the simulated population this translates to the constraint

$$\sum_{i=1}^{n} p_i x_i^1 \leq b_1.$$

If the population median of x^2 is known to be equal to b_2 then for the simulated population this becomes the constraint

$$\sum_{i=1}^{n} p_i u_i = 0.5$$

where $u_i = 1$ if $x_i^2 \leq b_2$ and it is zero otherwise. Hence, given a collection of population constraints based on prior information and a sample we will be able to represent the corresponding constraints on a simulated value of p by two systems of equations

$$A_{1,s} p = b_1 \tag{3.1}$$

$$A_{2,s} p \leq b_2 \tag{3.2}$$

where $A_{1,s}$ and $A_{2,s}$ are $m_1 \times n$ and $m_2 \times n$ matrices and b_1 and b_2 are vectors of the appropriate dimensions.

Let P denote the subset of the n dimensional simplex which is defined by equations (3.1) and (3.2). We assume the sample is such that P is non-empty and hence it is a non-full dimensional polytope. In this case the appropriate approximate version of the Polya posterior should just be the uniform distribution over P. We call this distribution the constrained Polya posterior (CPP). If one could generate independent observations from the CPP then one could find approximately the posterior expectation of population parameters of interest and find approximate 0.95 stepwise Bayes credible intervals. Unfortunately we do not know how to do this. Instead, one can use Markov chain Monte Carlo (MCMC) methods to find such estimates approximately. This can done in R (R Development Core Team 2005) and using the R package *polypost* which is available in CRAN. More details on the CPP and simulating from it are available in Lazar *et al.* (2008).

4 Constrained Polya posterior weights

A possible criticism of the Polya posterior and the CPP is that any simulated full copy of the population will only contain values of the characteristic that appeared in the sample. But it is exactly this property that will allow us to attach weights to the members of the sample.

We assume that we have a fixed sample for which the subset of the simplex defined by equations (3.1) and (3.2) is nonempty. For $j = 1, \ldots, n$ let

$$w_j = NE(p_j) = N\mu_j \tag{4.1}$$

where the expectation is taken with respect to the CPP. Note that the sum of the elements of $w = (w_1, \ldots, w_n)$ is the population size N and w_j can be thought of as the weight associated with the j^{th} member of the sample. These weights depend only on the observed values of the auxiliary variables and the known population constraints. Hence this is a stepwise Bayes method of attaching weights to the units in the sample which incorporates the prior information present in the auxiliary variables and does not depend explicitly on the sampling design.

We are assuming here that the population size N is know which may not always be the case. In such situations one could replace N in the above equation by an estimate. If the estimate is a good one then the resulting inferences for a population total should be satisfactory. When estimating a population mean the results would be much less sensitive to how close the estimate is to the true population size.

Much survey data which are used by social science researchers comes with weights attached to individual units. In such cases the CPP weights could be attached in the same way and the user would not need to use MCMC methods to calculate the weights. We will use the weights to define the Weighted Dirichlet posterior that can be used to find point and interval estimates of population quantities of interest at a relative modest computational cost. In the rest of the paper we will give examples to show that these weights can be used to generate inferential procedures with good frequentist properties.

But before proceeding we make a simple observation. Suppose we have in hand the sample along with a set of weights. If N is large, then we can construct a population where the proportion of units in the population of type (y_i, x_i) is w_i / N for $i = 1, \ldots, n$. Given the sample and the set of weights, we can think of this constructed population as the best guess for the unknown population. Then

$$\bar{y}_{bw} = \sum_{i=1}^{n} \frac{w_i}{N} y_i \text{ and } \sigma^2_{bw} = \sum_{i=1}^{n} \frac{w_i}{N} (y_i - \bar{y}_{bw})^2 \tag{4.2}$$

are the mean and variance of this constructed population.

5 The weighted Dirichlet posterior

It is often the case that weights are attached to data in public use files. These weights are then used by researchers to make point and interval estimates of population parameters. We shall see that the stepwise Bayes weights introduced here can often be used in standard frequentist formulas to estimate parameters of interest just as the usual weights are. We will use our weights to define the Weighted Dirichlet posterior (WDP) and show that it gives an alternative way to compute point and interval estimates for a variety of population quantities.

Let the w_j's be a set of weights defined by equation (4.1) with $\mu_j = w_j / N$. Consider the Dirichlet distribution over the simplex defined by the vector $n\mu = (n\mu_1, \ldots, n\mu_n)$ as an alternative posterior distribution for $p = (p_1, \ldots, p_n)$ when using the observed sample to generate complete simulated copies of the population. We will call this posterior the weighted Dirichlet posterior (WDP). Note the WDP is a looser version of the CPP. Under the CPP every complete copy of the population will satisfy the constraints; however, under the WDP, only the average of all the simulated populations will satisfy the constraints. It is easy to see that under the WDP

$$E\left(\sum_{i=1}^{n} p_i y_i \right) = \sum_{i=1}^{n} \mu_i y_i = \bar{y}_{bw} \tag{5.1}$$

and

$$
\begin{aligned}
V\left(\sum_{i=1}^{n} p_i y_i\right) &= \sum_{i=1}^{n} y_i^2 V(p_i) + \sum\sum_{i<j} y_i y_j \mathrm{Cov}(p_i, p_j) \\
&= \sum_{i=1}^{n} \frac{n\mu_i(n - n\mu_i)y_i^2}{n^2(n+1)} - 2\sum\sum_{i<j} \frac{n\mu_i n\mu_j y_i y_j}{n^2(n+1)} \\
&= \frac{1}{n+1}\left(\sum_{i=1}^{n} \mu_i(1-\mu_i)y_i^2 + 2\sum\sum_{i<j} \mu_i n\mu_j y_i y_j\right) \qquad (5.2) \\
&= \frac{1}{n+1}\left(\sum_{i=1}^{n} \mu_i y_i^2 - \sum_{i=1}^{n}\sum_{i=1}^{n} \mu_i\mu_j y_i y_j\right) \\
&= \frac{1}{n+1}\sigma_{bw}^2
\end{aligned}
$$

where $\bar{y}_{bw}$ and σ_{bw}^2 were defined in equation (4.2).

From this we see that when estimating the population mean, simulating from the WDP is equivalent to using the sample and their weights to construct the best guess for the population. In particular, when the weights are all equal the WDP is just the Polya posterior.

There are two main reasons for introducing the WDP. The first is that as the number of constraints used increases the approximate 0.95 credible intervals based on the CPP become too short and contain the true parameter value less than 95% of the time. This happens because with a large number of constraints the CPP does not allow enough variability in the simulated complete copies of the population which it generates. The second reason is that simulating from the WDP is much easier that simulating from the CPP. Now it would be possible to simulated from the constrained WDP in such a way that all the constraints would be satisfied but this involves as much effort as simulating from the CPP. Moreover, we believe that this would yield approximate 0.95 credible intervals which have poor frequentist coverage properties because they are too short.

Now suppose our set of weights is the reciprocals of the inclusion probabilities from the sampling design. Let $W = \sum_{i=1}^{n} w_i$. For most samples this value will not be equal to N but often is is quite close. Again we can construct our best guess for the population based on the weights. The mean and variance of this population will be

$$\bar{y}_{dw} = \sum_{i=1}^{n} \frac{w_i}{W} y_i \text{ and } \sigma^2_{dw} = \sum_{i=1}^{n} \frac{w_i}{W} (y_i - \bar{y}_{dw})^2. \tag{5.3}$$

If we use $\bar{y}_{dw}$ as an estimate of the unknown population mean then an unbiased estimate of its variance depends on the joint inclusion probabilities of the units in the sample. Since these are often difficult to obtain, what has been recommended in practice (Särndal, Swensson and Wretman 1992) is to assume the sampling was done with replacement even when that is not the case. Then the resulting approximate estimate of variance for $\bar{y}_{dw}$ is

$$\begin{aligned} \hat{V}_d(\bar{y}_{dw}) &= \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(n \frac{w_i}{W} y_i - \bar{y}_{dw} \right)^2 \\ &= \frac{\sigma^2_{dw} + \gamma_{dw}}{n-1} \end{aligned} \tag{5.4}$$

where the second line follows from some simple algebra and where

$$\gamma_{dw} = \sum_{i=1}^{n} \frac{w_i}{W} y_i^2 \left(n \frac{w_i}{W} - 1 \right). \tag{5.5}$$

Note that when the design is simple random sampling with or without replacement and $N = nk$ then $\gamma_{dw} = 0$. In this case, the estimate of variance in (5.4) is essentially equivalent to the variance in equation (5.2).

In situations where the Horvitz-Thompson estimator makes sense, calculations have shown that γ_{dw} tends to be negative. This suggests that in such situations intervals based on the WDP will tend to be conservative. However calculations also show that γ_{dw} term tends to be positive in situations where the Horvitz-Thompson estimator is not appropriate. We will see in such cases that the usual approximation can work poorly and intervals based on the WDP can have better frequentist properties.

6 Weights and Horvitz-Thompson

The usual definition of the weight assigned to a unit in the sample is the inverse of its inclusion probability. One is encouraged to think of a unit's weight as being the number of units

in the population which it represents. The resulting estimator of the population total is the Horvitz-Thompson (HT) estimator and is design unbiased. As we have already noted the unbiased estimate of its variance depends on the joint selection probabilities of the all the pairs of units appearing in the sample. Since in practice this can be impossible to compute the approximation in equation (5.4) is often used.

The HT estimator works best when y_i is approximately proportional to its selection probability. To compare its behavior to the WDP method we conducted a small simulation experiment. We constructed the variable x by drawing a random sample of 2,000 from a gamma distribution with shape parameter 5 and scale parameter 1 and adding 20 to each value. To generate y we let the conditional distribution of y_i given x_i be a normal distribution with mean $5x_i$ and standard deviation 20. The correlation of the resulting population was 0.49. We denoted this population by A. We created a second population, B, by using the same vector of x values but adding 400 to each y_i value. Our sampling plan used x to do sampling proportional to size, *i.e.*, pps(x). We used the R package *sampling* so that the inclusion probabilities were exact. Under this design we expect that the HT estimator would work well for population A but perform less well for population B. We also considered a third estimator, NHT, which is just the weights of the HT estimator rescaled so that they sum to the population size. We generated 500 samples of size 50. The results are giving in Table 6.1.

Table 6.1
Results for populations A and B based on 500 samples of size 50. The NHT estimator is the HT estimator renormalized so that the weights sum to the population size, N = 2,000. The nominal coverage for each method is 0.95.

Population	Method	Ave. abs err	Ave. len	Freq of coverage
A	HT	4,628	21,898	0.940
B	HT	8,965	43,914	0.960
A & B	WDP	4,706	24,381	0.960
A	NHT	5,051	21,897	0.896
B	NHT	5,051	43,919	0.998

Although not shown in the table both the HT and WDP estimators are unbiased for both populations. As expected the HT estimator is the best for population A although its performance falls off dramatically for population B. On the other hand the WDP performance for both populations is exactly the same. As a point estimator the NHT does much better than the HT estimator for population B but not as well for population A. Overall the WDP is clearly performs the best. What is an explanation for these differences?

In population A, $y_i \propto x_i$ and calculations show that γ_{dw} is almost always negative and its absolute value is small compared to σ_{dw}. In other words, when the HT estimator is appropriate it is essentially using the variance of the constructed population based on its weights to get its estimate of variance.

The only difference between populations A and B is that a constant has been added to the y value of each unit. Now if the sample weights allow us to make a good guess for the population in the first case what goes wrong in the in the second case to cause the HT estimator to preform so poorly? To see the problem consider the following.

In the HT estimate the sum of the weights in the sample almost never equal N, the population size. Given a sample in population B the HT estimate is

$$\sum_{i=1}^{50} w_i y_i = \sum_{i=1}^{50} w_i y_i' + 400 \sum_{i=1}^{50} w_i$$

where y_i' denotes the unit's corresponding value in population A and y_i its value in population B. Note the second term in the above equation is adding additional variablity to the HT estimator. In population B calculations show that the term γ_{dw} in equation (5.5) is positive and can be quite large. It is accounting for the extra variablity in the HT estimator in population B which results from that fact that here $y_i \propto x_i + 400$ and not x_i.

We note that Zheng and Little (2003) argued that when estimating a finite population total and when using a probability-proportional to size sampling design that a penalized spline, nonparametric, model based estimator generally outperformed the Horvitz-Thompson estimator.

Zheng and Little (2005) developed methods to estimate the variance of their estimator. Some related work can be found in Zheng and Little (2004).

The WDP weights only use the constraint that simulated complete copies of the population should have the correct population mean for x. This is a more robust assumption than the one which underlies the HT estimator. But to be fair to the HT estimator it should be remembered (as was pointed out by a referee) that it was developed with the limited goal of obtaining linear unbiased estimators of the population total. Today however its simplicity no longer seems so important when more complicated and efficient estimators are much easier to compute. The superior performance of the stepwise Bayes method here suggests that if one believes that they have a set of weights for the sampled units which sums to the population size and which yields a good guess for the population, then they should use the variance of their good guess for the population to construct an estimate of the variance of their estimate of the population mean rather than equation (5.4). This is particularly true for large surveys containing several y characteristics of interest. It would be very surprising if all of them satisfied the assumptions necessary to make equation (5.4) a good estimate of variance of a sample mean. Analogous to the observation in Royall and Cumberland (1981) and Royall and Cumberland (1985) that good balanced samples (the sample mean is close to the population mean) can lead to improved performance one should base their inference on simulated complete copies of the population which incorporate the available prior information contained in the auxiliary variables.

7 Examples

We believe that standard design based theory over emphasizes the role that the selection probabilities should play in making inferences after the sample has been observed. In this section we consider examples that show how the WDP can make use of objective prior information after the sample has been selected.

7.1 A simulation study

To further understand how using the stepwise Bayes weights in the WDP can work we did a simulation study. We constructed a population with 2,000 units and a single auxiliary variable, x. This variable was a random sample from a gamma distribution with shape parameter 5 and scale parameter 1. The conditional distribution of y_i given x_i was normal with mean $100 + (x_i - 8)^2$ and standard deviation 20. The correlation for the resulting population was -0.38. We denote this population by quad. Clearly this is a toy example and the particular form of the relationship between x and y is not important to the WDP methods beyond the fact that x does contain some information about y. In what follows we will compare WDP estimators to two standard methods under four different sampling plans.

To construct the CPP we assumed that the x values for the population are known and we use them to construct three strata after the sample has been observed. These strata will not be constructed in the usual way. We did this to underplay the usual role of the design and to emphasize the robustness of our approach against the choice of design. We will have a sample size of $n = 60$ and we will construct three post-strata. Let $x_{[1]} < x_{[2]} < \cdots < x_{[60]}$ be the order statistic of the x values in the sample. Let q_{20} and q_{40} be the population quantiles of $x_{[20]}$ and $x_{[40]}$ respectively. Then the CPP assumes that the total probability assigned to the units in the sample with the 20 smallest x values must be q_{20} and the total probability assigned to the next 20 smallest must be $q_{40} - q_{20}$. In other words we break the sample into three equal groups and use the information in the x values to get the appropriate population size of the corresponding strata. In addition the CPP assumes that the probabilities assigned to the sample must satisfy the population mean constraint for x.

The resulting WDP will be compared to two standard frequentist methods. The first is the post-stratified estimator which makes use of the same strata information as the CPP. The second is the usual regression estimator which assumes that the population mean of x is known. Although the regression estimator is not really appropriate for population quad it is included as a

comparison. When computing 95% confidence intervals for the population total both frequentist methods will assume simple random sampling even when different sampling designs were used. We will denote these two estimators by STR and REG respectively.

The first sampling design was simple random sampling without replacement. For the second we generated a set of sampling weights by taking a random sample of 2,000 from a gamma distribution with shape parameter 5 and scale parameter 1. We then added 5 to each value to get the vector, v say. Note the values of v and y are completely independent. We then used approximate $\text{pps}(v)$ where at each step the probability that a unit is selected is proportional to its v value and depends only the unselected units remaining in the population. We call this the Random Weights design. For the third design we used approximated $\text{pps}(x)$. For the fourth we found the linear function, say l, which maps the range of y onto the the interval $[1, 2]$. We then used approximate $\text{pps}(l(y))$ as the sampling design. We call this the y Dependent design. In this design the selection probabilities depend weakly on the y values and units with large y values are more likely to be selected than those with small values of y. In particular the unit with the largest y value is twice as likely to be selected as the unit with the smallest y value. Clearly the Random Weights design and the y Dependent design are not standard designs and would never be used in practice. They were included to emphasize our belief that in many cases given a sample a good estimate does not depend on how the sample was selected.

For each design we took 500 samples of size 60 and computed the point estimate, its absolute error, the length of its interval estimate and whether or not it contained the true parameter value. The results are given in Table 7.1.

Remember that in this example the WDP is using information from both the post-stratification and knowing the population mean of x while STR just uses the first and REG just uses the second. Under SRS and the Random Weights design all four methods preform about the same. For the other two designs WDP does the best. Over all four designs its frequency of coverage is closest to the nominal level of 0.95. Using the constraint involving the population mean of x

allows it to correct for some of the bias introduced by the sampling plans that STR cannot do. However this constraint can only do so much. If in the y dependent design the range of l was $[1, 4]$ then WDP's average absolute error is 4.5% better then that of STR and the frequency of coverage on the 0.95 nominal intervals were 0.86 and 0.80 respectively. There is just not enough information in x to correct for this much selection bias.

Table 7.1
Simulation results for population quad discussed in Section 7.1 for 500 random samples of size 60 for four different sampling plans. The true population total was 227,923.0. The nominal coverage for each method is 0.95.

Method	Ave. value	Ave. err	Ave. len	Freq of coverage
		SRS		
STR	227,856.1	4,165.0	21,332.1	0.950
REG	227,602.1	4,302.7	21,300.3	0.944
WDP	227,546.9	4,190.6	23,029.7	0.958
	Ave. min and max of WDP parameters were 0.658 and 1.580.			
		Random Weights		
STR	227,976.5	4,371.2	21,254.1	0.938
REG	227,715.5	4,462.2	21,305.9	0.934
WDP	227,721.2	4,420.6	22,901.4	0.950
	Ave. min and max of WDP parameters were 0.651 and 1.583.			
		pps(x)		
STR	225,295.8	5,228.9	23,008.4	0.916
REG	224,207.2	5,611.2	21,780.3	0.878
WDP	227,471.1	4,919.2	22,706.6	0.936
	Ave. min and max of WDP parameters were 0.374 and 3.024.			
		y Dependent		
STR	231,590.0	5,229.0	21,170.8	0.892
REG	231,424.4	5,143.4	21,127.9	0.902
WDP	231,139.1	4,967.6	22,867.0	0.938
	Ave. min and max of WDP parameters were 0.660 and 1.643.			

For each design we have included the average of the smallest and largest values of the parameter values defining the WDP which in this case must sum to 60. We see the range is largest for pps(x).

In the simulations we also used the WDP to construct 0.95 credible intervals for the population median of y. For the four designs its respective frequency of coverage was 0.956, 0.950, 0.952 and 0.930.

We did another simulation study where x was generated in the same way but now the conditional distribution of y_i given x_i was normal with $60 + x_i$ and standard deviation $2\sqrt{x_i}$. The correlation between x and y was 0.46. Under all four designs the performances of the point estimators were very similar. The WDP intervals tended to be a bit longer than the rest but over the four designs its average frequency of coverage for the population total was 0.949. Under the y Dependent design its frequency of coverage for the population total was 0.934 while for STR and REG the corresponding coverages were 0.896 and 0.886. Its average frequency of coverage for the population median of y was 0.942.

A frequentist could argue that this is an unfair example since the regression estimator does not make much sense for this population and of course they would be right. If for this problem you assumed a quadratic relationship between y and x and if you assumed that the first two population moments of x were known then the resulting regression estimator would out perform the WDP. In Lazar *et al.* (2008) there is such an example. Moreover, they show that including a constraint for the second moment of the CPP will hardly change the behavior of the resulting estimates. Hence, when there is good prior information about the model relating x and y this should be used in the analysis. When such prior information is not available we believe the WDP does have certain advantages even though it may not yield dramatic improvements over standard methods. It uses only objective prior information and makes no model assumptions about how the characteristics of interest and the auxiliary variables are related. It can correct for a slight dependency of the selection probabilities on the characteristic of interest. Although the sampling design plays no explicit role in its calculation, information which is often incorporated in the design can be reformulated as a constraint and be used when defining the CPP. Given a sample, inferences based on the WDP use many simulated complete copies of the population which on

the average are consistent with the prior information. This makes makes it straightforward to estimate parameters other than a population mean or total.

7.2 Stratification and estimating the median

In many applications only a few observations, sometimes only two, are taken from each stratum. For such problems finding a good confidence interval when estimating the population median can be difficult. Next we will compare the standard method, see for example Section 5.11 of Särndal *et al.* (1992), with the WDP. We will assume simple random sampling without replacement within strata.

For definiteness, assume we have L strata and stratum j contains N_j units. Let $N = \sum_{j=1}^{L} N_j$ be the total size of the population. Assume that two observations are taken from each stratum. Then the weight assigned to each sampled unit is one-half of the stratum size from which it was selected. The standard method uses these weights to find its confidence interval.

For this scenario the usual Polya posterior is applied within each stratum, independently across strata. Alternatively, this can be thought of as a CPP where the amount of probability assigned to the two sampled units in stratum j must sum to N_j / N. If $p_j = (p_{j,1}, p_{j,2})$ represents the probability assigned to the two sampled units from stratum j then under the CPP $E(p_j) = (N_j / (2N), N_j / (2N))$. Recalling the notation from Section 5 we see that under the WDP the weight assigned to each of the two sampled units in stratum j is $(LN_j) / N$. Recall that simulating complete copies of the population using the WDP means that individual simulated copies will almost certainly not satisfy the constraints however the constraints will be satisfied when we average over all simulated copies. At first glance this might seem like a bad idea but we will see that when estimating the population median interval estimates based on the WDP behave better than the standard intervals which are too short. We shall see that the extra variability present in the WDP yields longer intervals with better frequentist properties.

The stratified populations we considered were constructed as follows. The strata sizes were a random sample from a Poisson distribution with parameter $\lambda = 100$. The strata means were a random sample from a normal population with the mean $\mu = 150$ and with either a standard deviation of $\sigma = 10$ or $\sigma = 20$. The strata standard deviations were a random sample from a gamma distribution with scale parameter one and shape parameter $\gamma\sigma$ with either $\gamma = 0.10$ or $\gamma = 0.25$. We constructed two versions of each of the four types, one with 20 strata and the other with 40 strata. For each of the eight populations we took 500 samples where each sample consisted of two observations selected at random without replacement from each stratum. For each sample we compared the standard approach with estimates based on the WDP. The results can be found in Table 7.2. We only present the results for the 20 strata populations because the results for the 40 strata population are similar. Both methods are approximately unbiased and the point estimate based on the WDP seems to do just a bit better. But the confidence intervals produced by WDP are clearly superior. Even though in one case the WDP intervals are clearly too long its overall performance is much better than the standard intervals.

Table 7.2
Simulation results from 500 stratified random samples of size two within each strata from populations with 20 strata. The nominal coverage for each method is 0.95.

Method	Ave. value	Ave. err	Ave. len	Freq of coverage
		$\sigma = 10$ and $\gamma = 0.10$		
Stand	148.40	2.37	8.30	0.808
WDD	148.39	2.22	12.20	0.95
		$\sigma = 10$ and $\gamma = 0.25$		
Stand	144.28	5.70	20.59	0.834
WDD	144.18	5.41	28.38	0.950
		$\sigma = 20$ and $\gamma = 0.10$		
Stand	152.75	3.02	10.52	0.828
WDD	152.61	2.78	22.88	0.996
		$\sigma = 20$ and $\gamma = 0.25$		
Stand	155.94	6.72	23.17	0.826
WDD	155.89	6.35	34.96	0.962

What causes the poor performance of the WDP intervals in the one case? Additional simulations indicate that when the strata means vary widely and the strata variances tend to be relatively small then the WDP intervals will tend to be too long. In our simulations the case with $\sigma = 20$ and $\gamma = 0.10$ leads to a population with such strata. When the sample size was increased to four units per stratum the difference between the two methods is not so dramatic but the story remains much the same. The standard intervals tend to be to short and under cover while the WDP intervals are longer and tend to over cover.

Clearly the choice of a good method for constructing a confidence interval depends not only on the size of the intervals it produces and but on the probability with which those intervals fail to include the true but unknown parameter value. Cohen and Strawderman (1973) and Meeden and Vardeman (1985), among others, have explored the question of admissibility for confidence intervals. Although the results given there are not directly applicable to our case the second paper shows that in some situations certain Bayes procedures can yield almost admissible procedures. These type of arguments along with the fact that the standard interval is way too short gives some circumstantial evidence, we believe, that the WDP intervals in this example are not outrageously too long. To sum up, we believe that in the important special case when the sample sizes are two and the strata are not dramatically different the WDP intervals seem to be a serious competitor for the standard intervals.

7.3 Integrated public use microdata series

The Minnesota Population Center (MPC) is an interdepartmental demography research group at the University of Minnesota. A major goal of the MPC is to create databases and statistical tools which can be utilized in the study of economic and social behavior. One database of interest is the Integrated Public Use Microdata Series (IPUMS), which is a consolidation of U.S. censuses and other national surveys from 1850-present (Ruggles, Sobek, Alexander, Fitch, Goeken, Hall, King and Ronnander 2004). The word *microdata* is applied in this context because each row of

an IPUMS dataset corresponds to one individual or one household; such low-level of detail may be contrasted with a typical Census Bureau publication or online summary table, in which a preset geographic specific tabulation (geography can be the entire country, states, counties, census tracts *etc*.) of the microdata is given to the data user.

One dataset which offers a rich array of numerical variables is the 2005 American Community Survey (ACS). This Census Bureau product is a large sample survey, and the Census Bureau does not know the true population means for the variables. To conduct simulations with the 2005 ACS, the sample played the role of the population. More specifically, the full population was assumed to be a set of 3,579 Minneapolis residents who are of working age (between 25 and 75), and who earn a yearly wage between $20,000 and $120,000. For our purposes the two variables of interest were:

- *inctot*. Total pre-tax income from 2004.
- *sei*. The Duncan Socioeconomic Index. Created in the 1950's, this is a numerical variable which attempts to rate the prestige associated with an individual's occupation. The range of this variable is [1,100].

For our simulations we set $y = \log(inctot)$ and $x = sei$. The correlation between y and x is 0.398 and we assume that the mean of x is known. For estimating the population mean of y we considered the estimator based on the WDP and the regression estimator. We used two different designs: simple random sampling and approximate pps(x). In each case we took 300 samples of size 30. The results are given in Table 7.3. We see that although the two methods are comparable the WDP clearly gives the better intervals.

Table 7.3
Simulation results from 300 random samples of size 30 from the IPUMS population. The nominal coverage for each method is 0.95.

Design	Method	Ave. err	Ave. len/2	Freq of coverage
SRS	Reg	0.052	0.128	0.943
	WDP	0.052	0.138	0.947
pps(x)	Reg	0.062	0.132	0.897
	WDP	0.066	0.133	0.937

8 Final remarks

The construction of weights in survey sampling is often more of an art than a science. This is one possible conclusion that can be drawn from the recent paper of Gelman (2007) and the accompanying discussion. He argues for a Bayesian approach to constructing weights using regression models which relate the characteristic of interest to auxiliary variables. Here we argued for a stepwise Bayes approach which will make use of the information present in the auxiliary variables without assuming a model relating the characteristic of interest to the auxiliary variables. The resulting weight for a unit in the sample can be given the usual interpretation as the number of units in the population which it represents.

A frequentist weight, say w_i, is the inverse of an inclusion probability, and this number represents the number of units in the population represented by a particular unit in the sample. So $w_i \geq 1$ for all i and $\sum_{i \in s} w_i \approx N$. In Section 6 we saw that for the Horvitz-Thompson estimator the sum of the weights of the units usually fails to equal the population size which can result in a poor estimator except in very special circumstances. Another problem with frequentist weights is that they are often adjusted – after the sample is collected – to ensure that the frequentist estimates are in agreement with prior information about the population (Kostanich and Dippo 2002). After making adjustments, the weights may be rescaled so that they sum to a population total. However, the adjusted frequentist weights no longer depend just on the sampling design and they no longer represent inverses of inclusion probabilities. The intuition behind frequentist weights is therefore somewhat confusing. Before adjustments, frequentist weights are functions of the design; but after adjustments, they are now functions of the design and other prior information, which may or may not be related to the design.

Bayesians think of estimation in survey sampling as a prediction problem. Their predictions are based on an assumed model which can lead to weights being assigned to the units in the sample. See for example the aforementioned Gelman (2007) and Little (2004). As noted by a number of authors (Pfeffermann 1993) performing a weighted analysis for a model using inverses

of the inclusion probabilities can protect the sampler from model misspecification. Moreover in certain situations the two approaches may lead to similar results.

Recently, Rao and Wu (2010) have developed methods which use a pseudo empirical likelihood approach and base their inferences on Dirichlet posterior distributions. The resulting procedures, although formally somewhat similar to some discussed here, use prior information in a different way. For them much of the prior information must be filtered through the design while we believe that prior information which is often included in the design can be used directly to generate good posteriors. For better or worse we are closer to the classical Bayesian scenario where the posterior distribution does not depend on the sampling design.

Here we have focused on using the CPP to generated a set weights based on the sample and prior information and then making our inferences using the WDP based on these weights. Strief (2007) considered examples where the weights generated by the CPP were instead used in the appropriated frequentist formulas to get an estimate of variance and noted that their performance was similar to standard methods. Alternately one could imagine basing their inferences on the WDP but using frequentist weights, say generated by calibration methods (Särndal and Lundström 2005), instead. Although this deserves further study it is our expectation that such approaches should lead to inferential procedures with good frequentist properties.

In the design based approach consistency is an important property for an estimator to possess. For an important special case when the design is SRS the CPP estimators are consistent. This is demonstrated in Geyer and Meeden (2013).

Just as the CPP does, the WDP also has a stepwise Bayes justification. (For more details see Strief (2007).) The weights used in the WDP have a consistent formulation and interpretation. They are always a posterior expectation and always sum to the population size. They represent the average number of times that each unit in the sample appears in a simulated, completed copy of the population under the CPP. This average is with respect to the uniform distribution over all possible copies of the population which just contain the units in the sample and which satisfy the

given constraints. These weights depend only on the same kinds of objective prior information about the population which are often used to define and adjust frequentist weights. This allows them to incorporate prior information without explictly specifying a prior distribution.

In most cases the weight assigned to a unit in the sample will depend on the other units in the sample. We have argued that after the sample has been selected one should argue conditionally. That is, given the sample the weights should depend on all the available prior information about the population but not on how it was selected. (We are assuming that the person selecting the sample and the analyst are one in the same.) Any procedure constructed in this manner should preform well for a variety of sampling designs. For any procedure, be it either frequentist, Bayesian or stepwise Bayes this is the litmus test: it should be evaluated by how it behaves under repeated sampling from the design of interest.

To implement the methods discussed here one first needs to use the CPP to computed the weights for the observed sample. Then one needs to use the weights in the WDP to simulate complete copies of the population. The first step is the more difficult although the software package *polyapost* makes it relatively straightforward for anyone familiar with R. Once the weights are known it is easy to simulate from the WDP in many computer packages. This makes our approach more practical for survey datasets (like IPUMS) which are presented with the weights attached and are used by multiple researchers. A more serious limitation is that we have only considered simple single stage sampling designs. More work needs to be done to extend these methods to more complicated multi-stage designs. If the underlying constraints are selected wisely the resulting procedures can have good frequentist properties for a variety of sampling designs. These stepwise Bayes weights can be thought as our best guess for the unknown population given the sampled units and our prior information.

Acknowledgements

Research supported in part by NSF Grant DMS 0406169.

References

Binder, D. (1982). Non-parametric Bayesian models for samples from a finite population. *Journal of the Royal Statistical Society*, Series B, 44, 388-393.

Booth, J.G., Bulter, R.W. and Hall, P. (1994). Bootstrap methods for finite population sampling. *Journal of the American Statistical Association*, 89, 1282-1289.

Cohen, A., and Strawderman, W. (1973). Admissible confidence interval and point estimation for translation of scale parameters. *Annals of Statistics*, 1, 545-550.

Gelman, A. (2007). Struggles with survey weighting and regression modeling (with discussion). *Statistical Science*, 22, 153-188.

Geyer, C., and Meeden, G. (2013). Asymptotics for constrained Dirichlet distributions. *Bayesian Analysis*, 8, 89-110.

Ghosh, M., and Meeden, G. (1997). *Bayesian Methods for Finite Population Sampling*. Chapman and Hall, London.

Gross, S. (1980). Median estimation in survey sampling. In *Proceedings of the Survey Research Methods Section*, American Statistical Association, 181-184.

Hartley, H.O., and Rao, J.N.K. (1968). A new estimation theory for sample surveys. *Biometrika*, 55, 159-167.

Hu, M., Zhang, F., Cohen, M. and Salvucci, S. (1997). On the performance of replication-based variance estimation methods with small number of psus. In *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Kostanich, D.L., and Dippo, C.S. (2002). Design and methodology: 63rv. Technical report, The U.S. Census Bureau and The Department of Labor Statistics.

Lazar, R., Meeden, G. and Nelson, D. (2008). A noninformative Bayesian approach to finite population sampling using auxiliary variables. *Survey Methodology*, 34, 1, 51-64.

Little, R.J. (2004). To model or not to model? Competing modes of inference for finite poplation sampling. *Journal of the American Statistical Association*, 99, 546-556.

Lo, A. (1988). A Bayesian bootstrap for a finite population. *Annals of Statistics*, 16, 1684-1695.

Meeden, G., and Vardeman, S. (1985). Bayes and admissible set estimation. *Journal of the American Statistical Association*, 80, 465-471.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61, 317-337.

Rao, J.N.K., and Wu, C. (2010). Bayesian pseudo empirical likelihood intervals for complex surveys. *Journal of the Royal Statistical Society*, Series B, 72, 533-544.

R Development Core Team (2005). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, www.R-project.org.

Royall, R., and Cumberland, W. (1981). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, 71, 657-664.

Royall, R., and Cumberland, W. (1985). Conditional coverage properties of finite population confidence intervals. *Journal of the American Statistical Association*, 80, 355-359.

Rubin, D. (1981). The Bayesian bootstrap. *Annals of Statistics*, 9, 130-134.

Ruggles, S., Sobek, M., Alexander, T., Fitch, C.A., Goeken, R., Hall, P.K., King, M. and Ronnander, C. (2004). Integrated public use microdata series: Version 3.0 [machine-readable database]. University of Minnesota.

Särndal, C.-E., and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. New York: John Wiley & Sons, Inc.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer.

Strief, J. (2007). *Bayesian Sampling Weights: Toward a Practical Implementation of the Polya Posterior*. Ph.D. thesis, University of Minnesota.

Zheng, H., and Little, R. (2003). Penalized spline model-based estimation of finite population total from probability-proportional-to-size samples. *Journal of Official Statistics*, 19, 99-117.

Zheng, H., and Little, R. (2004). Penalized spline nonparametric mixed models for inference about a finite population mean from two-stage samples. *Survey Methodology*, 30, 2, 209-218.

Zheng, H., and Little, R. (2005). Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model. *Journal of Official Statistics*, 21, 1-20.


ELECTRONIC PUBLICATIONS AVAILABLE AT
PUBLICATIONS ÉLECTRONIQUES DISPONIBLE À
www.statcan.gc.ca

Survey Methodology, June 2013
Vol. 39, No. 1, pp. 29-58
Statistics Canada, Catalogue No. 12-001-X

Optimizing quality of response through adaptive survey designs

Barry Schouten, Melania Calinescu and Annemieke Luiten[1]

Abstract

In most surveys all sample units receive the same treatment and the same design features apply to all selected people and households. In this paper, it is explained how survey designs may be tailored to optimize quality given constraints on costs. Such designs are called adaptive survey designs. The basic ingredients of such designs are introduced, discussed and illustrated with various examples.

Key Words: Survey costs; Survey errors; Nonresponse; Responsive survey design.

1 Introduction

In most surveys, all sample units receive the same treatment and the same design features apply to all selected people and households. When auxiliary information is available from registry data or interviewer observations, then survey designs may be tailored to optimize response rates, to reduce nonresponse selectivity or more, generally, to improve quality. Although a general terminology is lacking in the literature, such designs are usually referred to as adaptive survey designs.

With this paper, we aim to describe the basic ingredients of adaptive survey designs, to systematize these designs by providing a mathematical framework, to illustrate their potential to improve efficiency of survey data collection, and to propagate their use in survey practice.

Adaptive survey designs assume that different people or households may receive different treatments. These treatments are defined before the survey starts, but may also be updated via data that are observed during data collection. In other words, allocation of treatments is based on data that are linked to the survey sample and on paradata. Paradata are data about the survey data collection process, *e.g.*, observations of interviewers about the neighborhood, the dwelling or the

1. Barry Schouten, Statistics Netherlands and University Utrecht, PO Box 24500, 2490HA Den Haag, The Netherlands. E-mail: jg.schouten@cbs.nl; Melania Calinescu, VU University Amsterdam, Department of Mathematics, De Boelelaan 1081, 1081HV Amsterdam. E-mail: melania.calinescu@vu.nl; Annemieke Luiten, Statistics Netherlands, PO Box 4481, 6401CZ Heerlen, The Netherlands. E-mail: a.luiten@cbs.nl.

respondents, or the performance of interviewers themselves. In this paper, paradata are used in the widest sense as data that are observed during data collection and that are informative about the response behavior of sampled people and households.

A general introduction to adaptive survey designs is given by Wagner (2008). Adaptive survey designs find their origin in the literature on medical statistics where treatments are varied beforehand over patient groups but also depend on the responses of patients, *i.e.*, depend on measurements during data collection. See for example Heyd and Carlin (1999), Murphy (2003) and Zajonc (2012).

A special case of an adaptive survey design is the responsive survey design. Responsive survey designs were introduced by Groves and Heeringa (2006). Like general adaptive survey designs, responsive survey designs may apply differential design features to sample units. However, the main distinction is that responsive survey designs identify promising and effective treatments or design features during data collection. In order to do so, the data collection is divided into multiple design phases. A new phase employs the outcomes of randomized contrasts between sample units in previous phases to distinguish effective from ineffective treatments and to identify costs associated with the treatments. Randomized contrasts are differences in response rates between subpopulations for randomly assigned design features. See for example Mohl and Laflamme (2007), Laflamme and Karaganis (2010), Phillips and Tabuchi (2009) and Peytchev, Riley, Rosen, Murphy and Lindblad (2010). The allocation of design features must be done in such a way that each phase reaches its phase capacity, which is the optimal trade-off between quality and costs. Responsive designs are motivated by survey settings where little is known about the sample beforehand and/or little information about the effectiveness of treatments is available from historic data. In these settings multiple phases are needed and responsive designs are practical. If the second and higher design phases of responsive designs are considered, however, then the starting point is similar to survey settings where substantial prior information about sample units is available or where a survey is repeated many times. The only distinction is

that in previous design phases part of the sample has already responded. In this paper, it is assumed that historic data are available, that effective treatments are identified beforehand and that it is specified what linked data and paradata are going to be used to adapt the design.

What is new in this paper? We make three contributions. First, we set up a general mathematical framework for optimizing response quality given cost constraints. Second, we explicitly allocate different design features to different sample units within this framework. Third, we propose to optimize quality indicators for nonresponse error. The last two contributions are by themselves not completely new. Simple adaptive survey designs are already applied, *e.g.*, in the Dutch Labour Force Survey larger households are not interviewed by web or telephone and proxy reporting is only allowed by a member of the household core. Attempts to optimize survey design accounting for nonresponse error go at least as far back as Hartley and Monroe (1979). And there is a vast literature on optimizing timing and number of contact calls in interviewer surveys, *e.g.*, Kulka and Weeks (1988), Greenberg and Stokes (1990) and Kalsbeek, Botman, Massey and Liu (1994). What is new is the ensemble of all the pieces into a general mathematical framework that abstracts from single design features and that allows to apply general quality indicators. The main motivations for the advance of such a framework are the strong pressure on survey costs and the rise of web as a survey mode. Web has a strong quality-cost differential; it is cheap but has low response rates and has different measurement properties than interviewer modes. As such, web challenges the trade off between quality and costs. Although survey literature has devoted considerable attention to trade-offs in survey designs between the various surveys errors, *e.g.*, Lyberg, Biemer, Collins, de Leeuw, Dippo, Schwarz and Trewin (1997) and Dillman (2007), in survey practice there are still surprisingly few cases where differential design features are investigated and implemented. With this paper, we hope to provide a steppingstone for future research and discussion into adaptive survey designs.

In Section 2, we describe theory and concepts behind adaptive survey designs. In Section 3, we present an example based on virtual survey data, and, in Section 4 we discuss a simulation study based on real survey data. Finally, in Section 5 we end with a summary and discussion.

2 What are adaptive survey designs?

2.1 Adaptive survey designs in general

In this section, a mathematical framework is set out for adaptive survey designs. In subsequent sections, components of this framework are highlighted and elaborated.

Let the population consist of units $k = 1, 2, \ldots, N$. The population of interest may consist of all units in a population but also of all recruited members of a panel. Each unit will be assigned a strategy s from the set of candidate strategies $S = \{\varphi, s_1, s_2, \ldots, s_M\}$. In the *survey strategy set* S the empty strategy φ is explicitly included. The empty strategy means that no action is undertaken, *i.e.*, the population unit is not sampled. This is the most general framework. In practice, one will often separate the sampling design from the strategy allocation and view the sample as given and fixed. However, one may include the decision to sample a unit explicitly in the overall allocation of resources.

In general a strategy s is a specified set of design features and may involve a sequence of treatments where treatments are only followed when all previous treatments failed. Some of those features may be sequential such as the type of contact mode and the type of survey mode, but the features may also describe different aspects of a survey design. Examples of strategies are

s_1 = (advance letter 1, web questionnaire, one reminder);
s_2 = (advance letter 1, web questionnaire, no reminder);
s_3 = (advance letter 2, CATI administered, maximum of six call attempts);
s_4 = (advance letter 2, CATI administered, maximum of 15 call attempts).

In the literature many design features are suggested and evaluated, *e.g.*, Groves and Couper (1998) and Groves, Dillman, Eltinge and Little (2002). We refer to De Leeuw (2008) for a discussion of survey modes, to Dillman (2007) and De Leeuw, Callegaro, Hox, Korendijk and

Lensvelt-Mulders (2007) for an elaboration of advance letters and reminders, to Wagner (2008) for a discussion of contact protocol, to Barón, Breunig, Cobb-Clark, Gørgens and Sartbayeva (2009) for a review of incentives, to Kersten and Bethlehem (1984), Cobben (2009) and Lynn (2003) for research into condensed questionnaires, to Moore (1988) for a discussion of proxy reporting, and to Cobben (2009) for an example of interviewer assignment.

It is assumed in this paper that the set of strategies S is known and fixed when strategy allocation is started. The set of strategies may be identified based on historical survey data, experience and pilot studies. We refer to Schouten, Luiten, Loosveldt, Beullens and Kleven (2010) and Schouten, Shlomo and Skinner (2011) for guidelines and examples on how to construct strategy sets.

With each population unit k a vector of covariates $X_k = (X_{k1}, X_{k2}, \ldots, X_{kp})^T$ is associated. X_k contains characteristics that are known before data collection starts and before strategies are allocated. The covariates must, therefore, be available in registrations or administrative data that can be linked to the sampling frame or in the sampling frame itself. Next to these general characteristics a second vector of covariates $\tilde{X}_k = (\tilde{X}_{k1}, \tilde{X}_{k2}, \ldots, \tilde{X}_{kq})^T$ may exist for unit k that reflects characteristics observed during data collection for sampled population units. These characteristics are termed paradata or process data because they are collected during the process of data collection by interviewers and data collection staff. However, other than the more traditional view on paradata as information about the process, in the adaptive survey design context $\tilde{X}_k$ contains observations about the sampled person or household. Examples of X_k are gender, age, type of household or educational level. Examples of $\tilde{X}_k$ are the interviewer assessment of the propensity to respond or the propensity to be contacted, the state of the dwelling or the neighborhood, and the presence of an intercom. $\tilde{X}_k$ is deliberately restricted to observations about the sample that allow for differentiation of survey design features. It does not contain the values of the design features themselves such as the interviewer that was assigned to the address.

The important distinction between X_k and $\tilde{X}_k$ is the level of availability. $\tilde{X}_k$ is known only for those units that are sampled and cannot be used in distinguishing subpopulations a priori. Let $q(x)$ represent the distribution of X_k in the population and $q(\tilde{x}, x)$ the joint distribution of X_k and $\tilde{X}_k$ in the sample. Furthermore, $q(\tilde{x} \mid x)$ denotes the conditional sample distribution. It is assumed that $q(x)$ and $q(\tilde{x}, x)$ are known in advance. In settings where no or little data can be linked, strategy allocation must be based fully on observations made during data collection.

Adaptive survey designs that allocate strategies based on population characteristics available in registry and frame data are termed *static*, while adaptive survey designs that allocate strategies that depend (also) on paradata are termed *dynamic*. It is important to remark that both static and dynamic designs have a strategy set that is fixed before data collection starts. However, for dynamic designs it is not known beforehand which strategies are going to be assigned to individual units because the choice of strategy depends on data that are observed during data collection.

Let $\rho(x, s)$ be the *response propensity* of a unit carrying characteristic $X = x$ and that is assigned strategy s. It is assumed that $\rho(x, s)$ is available from historic data, *i.e.*, from previous versions of the same survey, from surveys with similar topics and designs or from initial design phases. Obviously, the anticipated response propensity must be a close estimate of the true propensity. Section 2.4 returns to this essential component of adaptive survey designs.

The *expected costs* of the assignment of strategy s to a unit with $X = x$ is denoted as $c(x, s)$. It is an individual cost component. Literature tells us that survey costs consist of many components of which some are overhead and others are individual, *e.g.*, Groves (1989). Section 2.3 discusses cost functions.

Let $p(s \mid x)$ be the *allocation probability* of a population unit with characteristics x for strategy s, and let $p(s \mid x, \tilde{x})$ be the allocation probability to that strategy given that also paradata $\tilde{x}$ are observed. The following must hold

$$0 \leq p(s \mid x) \leq 1, \quad 0 \leq p(s \mid x, \tilde{x}) \leq 1 \tag{2.1}$$

$$\sum_s p(s \mid x) = 1, \quad \sum_s p(s \mid x, \tilde{x}) = 1, \tag{2.2}$$

i.e., all units are assigned a strategy. In general, allocation probabilities may have values between 0 and 1. In other words subpopulations with the same scores on x and $\tilde{x}$ may be (randomly) assigned to different strategies. For instance, only part of the non-respondents may be re-approached in a follow-up. Allowing for allocation probabilities between 0 and 1 increases the flexibility in meeting quality levels or cost constraints. In the following, p denotes the matrix of allocation probabilities, *i.e.*, $p = \{p(s_j \mid x, \tilde{x})\}_{1 \leq j \leq M, x, \tilde{x}}$ and contains the decision variables in the optimization.

The response propensities ρ_X can be derived from the strategy response propensities and the allocation probabilities by

$$\rho_X(x) = \sum_{s \in S} \sum_{\tilde{x}} q(\tilde{x} \mid x) p(s \mid x, \tilde{x}) \rho(s, x, \tilde{x}). \tag{2.3}$$

The strategies, covariates, response propensities, cost functions and allocation probabilities form the ingredients to adaptive survey designs. With these building blocks the adaptive survey design optimization problem can be formulated. Two ingredients are still missing, however, a quality function and an overall cost function. Let $Q(p)$ be some indicator of quality and $C(p)$ be an evaluation of total costs. The dependence on the allocation probabilities in both functions is stressed as the probabilities are the decision variables in the optimization.

The optimization problem can now be formulated as

$$\max_p Q(p) \text{ given that } C(p) \leq C_{\max} \tag{2.4}$$

or as

$$\min_p C(p) \text{ given that } Q(p) \geq Q_{\min}, \tag{2.5}$$

where $C_{\max}$ represents the budget for a survey and $Q_{\min}$ minimum quality constraints. Problems (2.4) and (2.5) are called dual optimization problems, although the solutions to both problems may be different depending on the quality and cost constraints.

It is important to stress that the optimization of quality or costs is done only once, before survey data collection starts, and is not repeated during data collection. Hence, it is the strategy that is adapted to the population unit, and in case of a dynamic design to paradata about that unit, but it is not the optimization itself that is adapted. The optimization is based on historic survey data that includes the paradata that has become available in a survey. The joint density function $q(x,\tilde{x})$, the response probabilities $\rho(x,\tilde{x},s)$ and the cost function $c(x,\tilde{x},s)$ are all estimated from historic survey data and are assumed to be given. Since in practice paradata becomes available only during data collection, the candidate strategies for units in the same stratum x are the same up to the moment the paradata $\tilde{x}$ becomes available. For instance, there may be the following four strategies: 1) two telephone call attempts and no follow-up, 2) two telephone call attempts and a follow-up with incentives, 3) three telephone call attempts and no follow-up, and 4) three telephone call attempts and a follow-up with incentives. The decision to make two or three call attempts is based on x, while the follow up is decided upon using a telephone paradata observation $\tilde{x}$. Thus, beforehand, it is estimated how many units will fall in stratum $(x,\tilde{x})$ and how many will receive a follow up, but only when $\tilde{x}$ is measured, the full strategy is known for individual units.

2.2 Quality objective functions

Adaptive survey designs, as discussed by the literature, typically focus on nonresponse error. In this section, we start with a general classification of quality functions, and then move to quality functions for nonresponse error. In general, a focus on nonresponse error is too narrow a view, especially, when the survey mode is one of the candidate design features in the adaptive survey design. Here, we do, however, not explicitly discuss other survey errors, but we return to

this issue in the discussion. We refer to Calinescu, Schouten and Bhulai (2012) for an extension of adaptive survey designs to measurement error and Beaumont and Haziza (2011) for a discussion on adaptive survey designs and nonresponse variance.

2.2.1 Covariate-based and item-based quality functions

When quality is optimized according to (2.4), then quality functions map the survey sample with linked data, paradata and answers to survey items to a single value which can be interpreted and optimized. When costs are minimized subject to constraints on quality as in (2.5), then quality may be multi-dimensional (but cost functions should be one-dimensional).

In general, two types of quality functions can be distinguished; quality functions that employ covariates from linked data and paradata only, and quality functions that also employ the answers to the survey target variables. We refer to them as *covariate-based* and *item-based*, respectively. An item-based quality function is a function of the response distribution of a survey item and the anticipated, estimated full population distribution given the available linked data and paradata. The main distinction between covariate-based and item-based quality functions is that item-based quality requires assumptions. Evidently, the answers of nonrespondents are missing. Hence, quality evaluation must be based on relations between target variables and covariates as observed in the response. As a consequence, there is a risk attached to item-based quality functions that originates directly from the phenomenon it attempts to measure. Relations between target variables and covariates may be different for nonrespondents and item-based quality may pose an incomplete image. Furthermore, in surveys with many survey target variables, different target variables may lead to different decisions and optimal survey designs. However, contrary to covariate-based quality functions, item-based quality functions tailor survey designs specifically to the topics of the survey. Covariate-based quality functions can only be related to the nonresponse bias of the covariates that are included.

2.2.2 Optimizing quality of response

We, first, describe briefly a number of quality functions that have appeared in recent literature. Next, we discuss the choice of a quality function.

The most well-known covariate-based quality function for nonresponse is the response rate. It is not a true covariate-based quality function in the sense that it depends on linked data or paradata. However, since the 0-1 response indicator may be viewed as the simplest form of paradata, it is termed a covariate-based quality function. The response rate is represented as the mean response propensity

Response rate: $$Q(p) = \bar{\rho} = \sum_{x,\tilde{x},s} q(x,\tilde{x})p(s \mid x,\tilde{x})\rho(x,\tilde{x},s) \tag{2.6}$$

Schouten, Cobben and Bethlehem (2009) propose two covariate-based quality functions, the R-indicator and a measure they call the maximal or worst-case nonresponse bias. The label of the second indicator is misleading as it is only an estimator of the maximal bias of the unadjusted mean of respondents, not the true maximal bias. A better label is the coefficient of variation of response propensities, which we will use here. The measures can be written as

R-indicator: $$Q(p) = R(\rho_Z) = 1 - 2S(\rho_Z) \tag{2.7}$$

Coefficient of variation: $$Q(p) = CV(\rho_Z) = \frac{S(\rho_Z)}{\bar{\rho}}, \tag{2.8}$$

where the representativeness may be evaluated with respect to linked data only, $Z = X$, or with respect to a vector containing both linked data and paradata, $Z = (X, \tilde{X})^T$. The standard deviations of the response propensities, $S(\rho_X)$ and $S(\rho_{X,\tilde{X}})$, can be written in terms of the strategy allocation probabilities as

$$S(\rho_X) = \sqrt{\sum_x q(x)\left(\sum_{\tilde{x},s} q(\tilde{x}\,|\,x)p(s\,|\,x,\tilde{x})\hat{\rho}(x,\tilde{x},s) - \bar{\rho}\right)^2} \tag{2.9a}$$

$$S(\rho_{X,\tilde{X}}) = \sqrt{\sum_{x,\tilde{x}} q(x,\tilde{x})\left(\sum_s p(s\,|\,x,\tilde{x})\hat{\rho}(x,\tilde{x},s) - \bar{\rho}\right)^2}. \tag{2.9b}$$

Särndal and Lundström (2010) and Särndal (2011a and b) propose indicators that are very similar in definition and nature to (2.7) and (2.8). These indicators were derived from the perspective of calibration and so-called balanced response, and may be used as alternatives to (2.7) or (2.8).

An example of an item-based quality function for nonresponse is presented by Groves and Heeringa (2006). For a specific target variable Y, Groves and Heeringa (2006) suggest the nonresponse bias of the unadjusted mean of respondents

Estimated nonresponse bias: $$Q(p) = \frac{\text{cov}(Y, \rho_X)}{\bar{\rho}}, \tag{2.10}$$

with $\text{cov}(Y, \rho_X)$ the response covariance between the target variable and the response propensities given covariates X. It can be written as

$$\text{cov}(Y, \rho_X) = \frac{\sum_x q(x)\rho_X(x)(\rho_X(x) - \bar{\rho}_X)(y(x) - \bar{y}_R)}{\bar{\rho}}, \tag{2.11}$$

with $\rho_X(x)$ as in (2.3), $y(x)$ the mean value of Y for $X = x$ and $\bar{y}_R$ the expected mean of respondents. Again, (2.11) can be extended to include paradata $\tilde{X}$.

All quality functions in this section are defined as population parameters. In practice, they need to be estimated from survey data. The true $\rho_X(x)$ need to be replaced by estimators $\hat{\rho}_X(x)$, based on some form of regression, and the summations over the population will be replaced by design weighted summations over the sample. We return to the estimation of propensities in Section 2.4.

Now, how to choose a quality function? All quality functions mentioned here attempt to measure the impact of nonresponse beyond that of a mere reduction in sample size. They do this based on covariates from linked data and paradata. The rationale behind optimizing these quality functions is that stronger traces of nonresponse error on these covariates may imply larger nonresponse error on other variables as well; the quality functions are viewed as process quality indicators rather than product quality indicators. Although appealing, this conjecture clearly

needs empirical support. The choice of a quality function for nonresponse should be based on the set of key survey variables, the population parameters of interest and the estimators that are going to be employed. The response rate and R-indicator do not aim at a specific population parameter or estimator. The coefficient of variation focuses on population means, but it is not specific to any survey variable or nonresponse adjustment, while the estimated nonresponse bias originates from the same perspective, but it is applied to a single survey variable. If a survey carries multiple key survey variables, then an item-based quality function for nonresponse is to be avoided, as it may lead to conflicting optimization problems. If a survey has a single key variable, then it is effective to either use an item-based quality function or to restrict to the most relevant covariates only in covariate-based quality functions. If a survey has multiple uses, then, in our view, it is too restrictive to focus on a specific population parameter and estimator, and we favor the R-indicator to any other quality function. If it can be expected that users will focus on population means or totals, then the coefficient of variation is to be preferred, in our opinion, as it does not assume a specific adjustment method.

However, even more important than the choice of the indicator is the set of linked data and paradata that are input to the indicator. If a survey has one or only a few key variables, then the selected linked data and paradata can and should relate to those variables. If, however, a survey has a wide range of survey variables, then one must restrict necessarily to auxiliary variables that generally distinguish persons or households.

So far, we have ignored the impact of nonresponse on precision, while requirements for the precision of the survey estimates may be given explicitly. There are two options to include precision in the optimization. First, one may add an additional constraint. The straightforward choice would be a constraint on the minimum number of respondents, possibly for a number of population subgroups; thereby avoiding to specify the population parameters and estimators. Second, the nonresponse variance may be included in the quality function itself, *i.e.*, one would consider indicators for the mean square error. This option is proposed and elaborated by

Beaumont and Haziza (2011). Under the second option, one again has to consider the set of key survey variables, the population parameter and the estimation strategy, as precision is specific to an estimator for some population parameter for a single survey variable.

2.3 Cost functions

Cost functions are the counterpart of the quality functions. There are several components to cost functions. It is important to restrict specification of costs relative to the design features that are varied in the adaptive survey design. For example, when it is the incentive that is differentiated with respect to different subpopulations, then costs need not be specified and detailed for interviewer traveling times. When it is the contact timing protocol that is the design feature that may be tailored, then, obviously, traveling times and traveling costs play a dominant role.

If a large number of design features is optional, then the cost functions have complex forms with many overhead and variable cost components. Generally, variable costs depend on the sample size while overhead costs do not. Overhead cost components may come from data collection staff, sampling and processing of samples. Variable costs arise for example from training and instruction of interviewers, mail and print of questionnaires and reminders, processing of paper questionnaires, interviewer hour rates and travel expenses, incentives and telephone number linkage, telephone usage and computer servers.

In the optimization two cost components may be identified: a fixed and a variable component. The variable component depends on the allocation of population units to strategies while the fixed component consists of all remaining costs. It must be stressed that the fixed component is different for adaptive survey designs that focus on different design features. The cost function $C(p)$ is the sum of two components

$$C(p) = C_F + C_V(p), \tag{2.12}$$

of which only the second, the variable component, depends on the allocation probabilities.

In general, with a survey strategy s, costs $c(\tilde{x}, x, s)$ are associated with population units from group $(\tilde{x}, x)$. The individual cost function may be a function of response propensities, or even more specifically of contact and participation propensities. For instance, the interviewer costs in different contact timing protocols depend on the contact rates of the selected subpopulations. The cost function $c(\tilde{x}, x, s)$ is a relative cost function as it describes only the contribution of the strategy to the variable cost component $C_V(p)$

$$C_V(p) = \sum_{\tilde{x},x,s} q(\tilde{x}, x)p(s \mid x, \tilde{x})c(\tilde{x}, x, s). \tag{2.13}$$

Three remarks are in order. First, the derivation of fixed and variable cost components is complicated when a survey organization runs many surveys in parallel. On the one hand, the interaction between surveys makes it hard to separate costs per survey, especially when strategies are tailored. On the other hand, when multiple surveys are conducted some of the variable costs components may be labeled as fixed. For example, when only a relatively small number of population units are assigned to the face-to-face survey mode, then traveling costs may be assumed to remain unchanged as the addresses are clustered with addresses from other surveys. The second remark concerns the multidimensional aspect of costs. Apart from the overall budget it may be requested that interviewer occupation rates are close to one throughout time or that none of the interviewers has to work overtime more than a fixed amount of time. As a consequence, the cost function becomes a vector and the constraint a vector of constraints. The third remark concerns the validity of the cost functions. Since cost functions are hard to construct in practice, it may turn out that the optimization was too optimistic. It is important to monitor data collection closely and to build in indicators for strategies.

2.4 Estimating response probabilities

Next to cost parameters and quality functions, the other important ingredient of adaptive survey designs is the set of response propensities for the various strategies. Such propensities need to be known from past surveys, preferably the same survey or otherwise a similar survey.

Alternatively, as Groves and Heeringa (2006) propose, one may use earlier phases of the data collection to learn and derive propensities. This will be at the expense of efficiency since part of the survey is already conducted. Nonetheless, the gathered information directly feeds back to the current survey.

Literature on household surveys gives an extensive list of models for response that include design features. The common denominator in all models is that response propensities are estimated based on a number of assumptions about the true nature of the nonresponse missing-data mechanism. In general such models are simplifications. Consequently, anticipated response propensities $\rho(x,s)$ have a standard error, and may even be biased themselves when they are based on similar, but different surveys. In the optimization, this uncertainty can be accounted for by allowing response propensities to be random variables rather than fixed quantities. The randomness demands for sensitivity analyses and evaluations of the robustness of the optimization that provide insight into the variation of quality and costs when the survey is conducted multiple times (under the same circumstances).

2.5 The optimization problem

One may take two approaches to the optimization of (2.3) and (2.4): a trial-and-error approach or a mathematical optimization. In this paper, we concentrate on a mathematical framework and optimization, but one may be more modest and introduce adaptive survey designs gradually through pilot studies and field tests.

Quality functions (2.6), (2.7), (2.8) and (2.10) all are functions of the strategy allocation probabilities p. The response rate is a linear function of the allocation probabilities, which makes it relatively easy to optimize using standard optimization software (*e.g.*, the *linprog* package in R or any other software that can address linear programming problems). Still, as far as we know, due to the high dimensionality of p there is no closed form solution to (2.4) even for linear problems. In general, however, the quality functions are nonlinear, nonconvex functions with

respect to the allocation probabilities, and cannot be optimized without numerical or Monte Carlo methods. The complexity of the problem grows quickly as a function of the number of candidate strategies and the number of subgroups based on linked data and paradata.

Current statistical softwares contain procedures or packages that can handle nonlinear optimization problems, like *nlm* or *nlminb* in R or *proc optmodel* in SAS. However, nonlinear nonconvex problems may require long computational times or may converge to local optima. For this reason, specialized optimization softwares such as Xpress, Baron or AMPL are recommended.

In the examples of Section 3 and 4, we perform a number of optimizations. The optimization problem of Section 3 is relatively simple; the quality objective function is the R-indicator which is evaluated against two population subgroups. For two subgroups the optimization can be rewritten as a linear programming problem. For the example of Section 4, we were able to construct an algorithm that converges to the optimal solution in a small number of steps. All optimizations were programmed in R and the code is available upon request.

3 A dynamic adaptive survey design: Re-assigning interviewers in a follow-up survey

In this section, we provide an example of a dynamic adaptive design: the re-assignment of interviewers based on observations of the propensity to cooperate. The example is based on hypothetical response propensities and cost functions. Interviewers are assigned to sample cases that have refused once, based on an assessment of the propensity to respond made during a first phase of the survey. The assessment is made for respondents and refusers, but it is not available for sample units who were not contacted during the first phase. It provides a judgement on the propensity that the sample unit participates in the survey when contacted again. The assessment is made on a three point scale: *easy, medium, difficult*. Easy means that there is a high probability that if contacted again the sample unit would respond.

After a first phase of data collection, the intermediate survey results are evaluated and sample units are divided into respondents, refusers and noncontacts. Refusers receive a different treatment. Interviewers are rated based on their historic performance and grouped in *good* and *less good* interviewers. Refusers are re-assigned to one of the two groups of interviewers. Since there is no assessment available for non-contacts, the treatment for this group is not altered.

We use the R-indicator given by (2.7) as the quality objective function. We split the sample using $X = (\text{age})$ into two groups, labelled as *young* and *old*. The goal in the second phase is to assign refusers to the two interviewer groups such that the R-indicator with respect to age is maximized.

Let n be the sample size of the survey. The population proportions of the two subpopulations, *young* and *old*, are denoted by $q(1)$ and $q(2)$. We let $q(\tilde{x}\,|\,x)$ be the conditional probability that a sample unit from age subpopulation x is of type $\tilde{x}$, where $\tilde{x} \in \{\text{easy, medium, difficult}\}$. Furthermore, let $\lambda(x, \tilde{x})$ be the probability that a sample unit of type $\tilde{x}$ from age subpopulation x is a refusal. If a person is not a refuser, then $\mu(x, \tilde{x})$ is the probability that the person either was a respondent after the first phase or becomes a respondent when he/she was a noncontact after the first phase.

The total number of interviewers is M and $p_s M$ represents the number of interviewers with skill $s \in S = \{\text{good, less good}\}$, $0 \le p_s \le 1$ and $p_{\text{good}} + p_{\text{less good}} = 1$. The set S forms the set of strategies, *i.e.*, we want to assign each refuser to either a good or a less good interviewer. We assume that each interviewer can handle at most c refusal cases in the second phase of the survey. The probability that a refusal of type $\tilde{x}$ from subpopulation x will respond if contacted by an interviewer of skill s is denoted by $\rho(s, x, \tilde{x})$ and it is again assumed to be known from previous surveys.

Let $\{p(s\,|\,x, \tilde{x})\}_{x,\tilde{x}}$ be the set of decision variables, where $p(s\,|\,x, \tilde{x})$ represents the probability that a sample unit of type $\tilde{x}$ will be assigned to an interviewer of skill s given that he/she

belongs to subpopulation x. In other words, we allow for a random assignment of sample units to the two interviewer groups.

In this example, we express costs in terms of the overall interviewer occupation rates. Since interviewers can handle at most c cases, there are two constraints

$$n\sum_{x,\tilde{x}} q(x)q(\tilde{x}\,|\,x)\,p(s\,|\,x,\tilde{x})\,\lambda(x,\tilde{x}) \leq Mp_s c, \quad \forall s \in S.$$

In other words, the total number of refusers that can be assigned to interviewers of skill s is restrained to the maximum possible workload for that skill group.

The response propensity for a unit from subpopulation x can now be derived as

$$\sum_{\tilde{x}} q(\tilde{x}\,|\,x)\Big[\,(1-\lambda(x,\tilde{x}))\mu(x,\tilde{x}) + \lambda(x,\tilde{x})\sum\nolimits_s p(s\,|\,x,\tilde{x})\rho(s,x,\tilde{x})\,\Big],$$

and form the input to the R-indicator.

Now, consider the following input data for the example: a sample size of $n = 2{,}000$, a total of 80 interviewers, $M = 80$, a maximal workload of 30 cases per interviewer, $c = 30$, an age distribution equal to $q(1) = q(2) = 0.5$, conditional distributions of refusal type $q(\tilde{x}\,|\,1) = (0.2, 0.3, 0.5)'$ and $q(\tilde{x}\,|\,2) = (1/3, 1/3, 1/3)'$ and 25% of the interviewers are classified as good, $p_1 = 0.25 = 1 - p_2$.

Tables 3.1 and 3.2 give the hypothetical response probabilities $\rho(s,x,\tilde{x})$ for the two subgroups when refusal conversion is applied, as well as the cooperation probabilities $\mu(x,\tilde{x})$ and refusal probabilities $\lambda(x,\tilde{x})$.

We optimize the R-indicator with respect to the two age groups. For two strata, it can be shown that the R-indicator is maximal when the absolute distance between the two strata response propensities is minimal. The optimal value of the R-indicator turns out to be 0.827. Table 3.3 shows the optimal values of the decision variables; all but one of the decision variables $p(s\,|\,x,\tilde{x})$ are either 0 or 1, *i.e.*, the re-assignments are mostly non-probabilistic. The exception is the subpopulation of young persons with medium response propensity assessment.

Table 3.1
Response probabilities when refusal conversion is applied to young and old refusers given the assessment of propensity to respond.

	Young refuser					
	Good interviewer			Less good interviewer		
	Easy	Medium	Difficult	Easy	Medium	Difficult
$\rho(s, 1, \tilde{x})$	0.8	0.6	0.4	0.7	0.5	0.3
	Old refuser					
	Good interviewer			Less good interviewer		
	Easy	Medium	Difficult	Easy	Medium	Difficult
$\rho(s, 2, \tilde{x})$	0.9	0.7	0.5	0.8	0.6	0.4

Table 3.2
Refusal and cooperation probabilities in the first phase of data collection

	Young			Old		
	Easy	Medium	Difficult	Easy	Medium	Difficult
$\lambda(x, \tilde{x})$	0.5	0.6	0.7	0.2	0.3	0.4
$\mu(x, \tilde{x})$	0.85	0.8	0.76	0.95	0.93	0.91

Table 3.3
Optimal assignment of cases to interviewers

	Young			Old		
	Easy	Medium	Difficult	Easy	Medium	Difficult
Good	1	0.83	1	0	0	0
Less good	0	0.17	0	1	1	1

It is useful to compare the optimal allocation to a random allocation of interviewers in order to see how much is gained. If we would randomly assign the refusals to the interviewers, then the value of the R-indicator equals 0.749. The optimal assignment, thus, leads to a considerable

increase in the R-indicator. The response rates are, respectively, 72.0% and 70.1% for the optimal and the random assignment.

If we increase the number of interviewers, while fixing the maximal number of cases per interviewer as well as the other parameters, then for any interviewer number higher than $M = 84$ the R-indicator does not improve. Both interviewer groups are sufficiently big to handle the entire sample and the cost constraint is no real constraint anymore. The R-indicator for $M = 84$ is equal to 0.830 and the response rate is 72.1%. If we would maximize the response rate rather than the R-indicator, then the allocation of interviewers will converge towards assigning only *good* interviewers to all cases.

4 A static adaptive survey design: Assigning telephone interviewers

In this section, a simulation study is presented where telephone interviewer assignment is the design feature of interest. The response probabilities used in the example are estimated from real telephone survey data.

The Dutch Survey of Consumer Satisfaction (SCS) is a monthly telephone survey about the sentiments of households about their economic situation and expenditure. The survey provides insight into short-term economic development, and early indicators of differences in consumer trends. Each month 1,500 households are sampled. The two most influential causes of nonresponse in the SCS are non-contact and refusal. Of the sample 95% is contacted, and of the contacted 71% of the households participate. The response rate is 67%.

One of the most important factors that affect participation is the interviewer. Interviewer's performance may vary greatly when it comes to obtaining response. In total 60 interviewers worked on the SCS during 2005. That means an interviewer had contact with 280 households on average. Interviewer participation rates ranged from 50% to 79%. The lowest rate of 50% was, however, exceptional as the one but lowest participation rate was 61%. The mean interviewer

participation rate was 67%. Households were randomly assigned to interviewers in the CATI management system. Hence, with respect to the interviewer the data are randomized (or interpenetrated). In the following, the interviewer will be the design feature of interest. The survey strategy set S consists of sixty strategies, $S = \{s_1, s_2, \ldots, s_{60}\}$.

From the available auxiliary variables a vector X was selected containing ethnicity, gender composition of the household core (male, female or mix), average age of the household core in 5-year classes, type of household, degree of urbanization of the neigborhood of residence and average value of houses in the neighborhood. Especially age, average house value and type of household relate to key statistics deduced from the SCS. No paradata were available in this study. Therefore, the adaptive survey design is static. In the optimization the allocation probabilities $p(s_k \mid x)$ need to be chosen, *i.e.*, it needs to be decided to which interviewers subpopulations based on X are assigned $\left(\text{such that } \sum_k p(s_k \mid x) = 1\right)$.

The coefficient of variation of the response propensities ρ_X defined by (2.8) is selected as the target quality function. To estimate the response propensities $\rho(s_k, x)$ for interviewers, a multilevel model is used with the identity link function, *i.e.*, a linear regression with two levels. The interviewers form the first level of the model and the households the second level. The multilevel model is used to separate individual response propensities and interviewer response propensities. The rationale is that by separating interviewer and individual, the interviewer effect can be isolated and interviewer assignment can be optimized. We chose a linear model as it allows for easy optimization. Since the propensities are never close to 0 or 1, the linear model produces almost the same estimates as a logit or probit model.

For the interviewer effect it was first investigated whether it was sufficient to use a fixed slope multilevel model, *i.e.*, the interviewer is added as a main effect only and there are no interactions with auxiliary variables. All pre-selected covariates gave significant contributions to the multilevel model, but none of the interactions with the interviewer were significant at the 5% level. For this reason, we restrained ourselves to the following main effect model

$$\rho(s_k, x_i) = \beta_0 + \beta_X x_i + \beta_k \tag{4.1}$$

where x_i is the covariate vector of household $i, 1 \leq i \leq n, n$ the sample size, β_k is the (fixed) interviewer effect for interviewer k, β_0 is the constant term or intercept and β_X is the slope parameter. We let $\rho(x_i) = \sum_k p(s_k \mid x_i)\rho(s_k, x_i)$ denote the response propensity of sample unit i.

Model (4.1) was fitted to the SCS data set. Next, the estimated interviewer effect β_k was used to optimize the coefficient of variation, subject to two cost constraints: both the total interview time and the individual number of calls for each interviewer must be the same as in the original design. Since the telephone management system handles the calls, the interview time is the dominant component in the costs. If we fix the total interview time, then we constrain costs to be the same as for the regular SCS. Since interviewers can handle only a certain amount of calls, we must also fix the number of calls they are allocated to. The first constraint implies that we fix the response rate, as the total interview time is the multiple of the average individual interview time and the number of respondents. The SCS questionnaire is simple and does not contain any nested sets of survey items. As a result the individual interview time shows hardly any variation over population subgroups. The second constraint is equal to

$$\sum_i p(s_k \mid x_i) = n_k, \tag{4.2}$$

where n_k is the pre-specified number of calls for interviewer k and $\sum_k n_k = n$.

We optimize the coefficient of variation by distributing the β_k's to the households. Due to the additive nature of the model, it is easy to show that any permutation of the interviewers to the cases leads to the same average response propensity and, hence, to the same interview time and costs. The average response propensity is

$$\bar{\rho} = \frac{1}{n}\sum_i \rho(x_i) = \frac{1}{n}\sum_{i,k} p(s_k \mid x_i)(\beta_0 + \beta_X x_i + \beta_k) = \beta_0 + \frac{1}{n}\sum_i \beta_X x_i + \frac{1}{n}\sum_k n_k \beta_k,$$

which does not depend on the set of allocation probabilities $p(s_k \mid x)$. As a consequence, optimizing the coefficient of variation amounts to optimizing the variance of the response propensities $S^2(\rho_X)$.

If we restrict ourselves to 0-1 decision variables, *i.e.*, $p(s_k \mid x) \in \{0,1\}, \forall x, k,$ then it is relatively easy to show that the optimal allocation corresponds to linking the best interviewers to the most difficult sample units and vice versa. In other words, the sample units are sorted by putting the individual response propensities without the interviewer effect, $\beta_0 + \beta_X x_i,$ in an increasing order, and the interviewers are sorted in a decreasing order based on their interviewer effect, β_k. If two sample units i and j are allocated to two different interviewers, say k and l, and $\beta_X x_i < \beta_X x_j, \beta_k < \beta_l$ and $p(s_k \mid x_i) = p(s_l \mid x_j) = 1,$ then it is optimal to switch the two interviewers, *i.e.*, $p(s_l \mid x_i) = p(s_k \mid x_j) = 1.$ This can be shown as follows. The difference in variance $S^2(\rho_X)$ is proportional to

$$\begin{aligned}\Delta S^2(\rho_X) &= (\beta_0 + \beta_X x_i + \beta_k - \bar{\rho})^2 \\ &\quad +(\beta_0 + \beta_X x_j + \beta_l - \bar{\rho})^2 - (\beta_0 + \beta_X x_i + \beta_l - \bar{\rho})^2 - (\beta_0 + \beta_X x_j + \beta_l - \bar{\rho})^2 \\ &= 2(\beta_l - \beta_k)(\beta_0 + \beta_X x_j - \bar{\rho}) - 2(\beta_l - \beta_k)(\beta_0 + \beta_X x_i - \bar{\rho}) \\ &= 2(\beta_l - \beta_k)(\beta_X x_j - \beta_X x_i) > 0.\end{aligned} \tag{4.3}$$

From (4.3), we can conclude that there is a decrease in variance, and, hence, in the coefficient of variation, if we swap the two interviewers for cases i and j. From this argument, it follows easily that the optimal solution is as suggested. In a similar fashion, but requiring more algebra, it can be shown that the optimal solution for probabilistic allocations, $p(s_k \mid x) \in [0,1],$ is the same.

The first two rows of table 4.1 contain the average response propensity and the coefficient of variation before and after re-assignment of interviewers. The coefficient of variation dropped from 0.117 to 0.035. In order to get an idea of the significance of the change in the quality function, we computed bootstrap standard errors. For each bootstrap, the re-assignment of interviewers was performed. The errors are given in table 4.1.

Table 4.1
The average response propensity and coefficient of variation of the regular SCS, the SCS after re-assignment of interviewers without and with adjustment for interview time. Bootstrap standard errors are given within brackets.

SCS	Adjustment for interview time?	$\bar{\rho}$	$Q(p)$
Regular	-	70.8%	0.117 (0.005)
Re-assignment	No	70.8%	0.035 (0.003)
Re-assignment	Yes	70.8%	0.034 (0.003)

The reader may have noticed that fixed numbers of interviewer cases do not imply fixed numbers of interviews per interviewer. In fact, by rearranging the interviewers, the good interviewers will do fewer interviews as they get the harder cases, while the less good interviewers do more interviews. As a result, the good interviewers will work smaller numbers of hours than they would do in the regular SCS and the less good interviewers will work more. This would be an undesirable side effect, which can, however, be adjusted relatively easy. Starting from the optimal solution, and sorting again the sample units based on their individual response propensities without the interviewer effect, we can shift neighbouring cases from less good interviewers to better interviewers. This is done in such a way that the total interview time per interviewer does not exceed that of the regular SCS. One can again prove that this procedure leads to a new optimal solution where the constraint on the fixed number of cases in (4.2) is replaced by the constraint on the fixed number of interviews

$$\sum_i p(s_k \mid x_i)\,\rho(s_k, x_i) = r_k, \tag{4.4}$$

where r_k is the pre-specified number of interviews. Table 4.1 presents the coefficient of variation for the optimal solution given (4.4). The response rate remains fixed, and the coefficient of variation is marginally smaller.

In 2009, the SCS survey has been used as an instrument to test a static adaptive survey design. We refer to Luiten and Wetzels (2009) and Luiten and Schouten (2013) for details. Interviewer

assignment was one of the main design features that were adapted. Other design features were the survey mode and the contact protocol. Apart from telephone, also web was selected as a potential survey mode. Sample units with low estimated contact probabilities were assigned to more intensive contact protocols and were prioritized. Based on historical SCS data, contact and response probabilities were estimated. The pilot succeeded in significantly improving the coefficient of variation, while fixing the response rate and budget.

In this section, we presented a simulation study where good telephone interviewers get more difficult cases. This may in practice lead to annoyance among these interviewers. When implementing such a design, one should carefully instruct interviewers beforehand. In the 2009 SCS pilot, this did not lead to any negative comments from interviewers. In face-to-face surveys, a re-assignment of interviewers cannot be done so easily as travel costs may change drastically. Still, within densely populated inteviewer regions, re-assignment may be an option.

5 Discussion

This paper describes survey designs in which different population units receive different treatments or survey strategies. Differences between population units are reflected by covariates from either linked data from registrations or paradata. Survey strategies are defined as different specifications of survey design features. Such designs are termed adaptive survey designs as they adapt or tailor data collection to the population of interest. Basic ingredients of adaptive survey designs are survey strategies, population covariates, response propensities, cost and quality functions and strategy allocation probabilities. Adaptive survey designs attempt to optimize response quality by assigning different strategies to different population units. The strategy allocation probabilities represent the decision variables in the optimization.

We believe this paper contributes to the literature in three ways: it presents a general framework, it explicitly opts to choose from a set of strategies in making a quality-cost trade off,

and it focuses on indicators for nonresponse error. The last two components can be found in the survey literature; it is the generalization to multiple design features and nonresponse error that is new. In its most modest form, adaptive survey designs are a stratified allocation of survey strategies over different population subgroups. In its most ambitious form, adaptive survey designs are extensions of sampling designs to multiple strategies and with a focus on nonresponse error. However, even in its most modest form, adaptive survey designs may include survey modes, incentives, reminders, length of fieldwork in face-to-face surveys, interviewer assignment and type of reporting.

Adaptive survey designs lend themselves best to settings where surveys are run repeatedly for a longer time period. In such settings, the historic information is strongest. The designs also lend themselves to survey institutes that conduct many surveys that are relatively similar in topics and budget. New and one-time only surveys ask for modesty and caution. However, this would also be true for single strategy designs. Adaptive survey designs may account for the lack of strong historic data by allowing for uncertainty in response propensities and other parameters, and by introducing a learning period or initial design phase.

In our view, the focus on nonresponse error is an important part of the framework. In this paper, we aim at representativeness of response. This aim comes from our conviction that nonresponse is always not-missing-at-random. We see larger deviations from missing-completely-at-random mechanisms for relevant auxiliary variables as indications of stronger not-missing-at-random nonresponse on survey variables given these auxiliary variables. Theoretically, this does not have to be true. Consider a simple binary yes-no survey question and 50% nonresponse. The extreme cases arise when all nonrespondents would say either yes or no. They can do so regardless of the choice of auxiliary variables and, hence, the maximal nonresponse bias on this question is the same for whatever choice of auxiliary variables. Hence, research should provide empirical support for the focus on indirect measures for nonresponse error.

Future research into adaptive and responsive designs is also needed for other questions. Research should extend designs to multiple survey errors and should investigate the robustness of designs for misspecification of models for response propensities. Until now, adaptive and responsive survey designs have focused on the nonresponse error and ignored the response or measurement error. It is well known, however, that some survey design features, *e.g.*, survey mode or interviewers, may have a strong impact on the response error and, consequently, on the total survey error. Adaptive survey designs should, therefore, account for measurement error as well, when it can be expected that design features have a strong differential impact on response error. Optimization accounting for multiple errors represents an important area of future research.

Adaptive survey designs should in all cases be modest in the number of strategies employed in order to avoid an overly complex survey process and optimization on propensities and cost functions that are subject to uncertainty. Nevertheless, a structured way of looking is always to be preferred; adaptive designs provide such a framework and accommodate a structured search for enhanced survey designs.

Acknowledgements

The authors thank James Wagner, Mick Couper, Fannie Cobben and Mariëtte Vosmer for their useful comments on the draft version of this paper. The authors also thank the associate editor and referees for their useful remarks which have greatly improved the paper.

References

Barón, J.D., Breunig, R.V., Cobb-Clark, D., Gørgens, T. and Sartbayeva, A. (2009). Does the effect of incentive payments on survey response rates differ by income support history. *Journal of Official Statistics*, 25, 483-507.

Beaumont, J.-F., and Haziza, D. (2011). A theoretical framework for adaptive collection designs. Paper presented at 5th International Total Survey Error Workshop, June 21-23, Quebec, Canada.

Calinescu, M., Schouten, B. and Bhulai, S. (2012). Adaptive survey designs that minimize nonresponse and measurement risk, Discussion paper, Statistics Netherlands. Available at through www.cbs.nl.

Cobben, F. (2009). Nonresponse in sample surveys. Methods for analysis and adjustment, Ph.D. Thesis, University of Amsterdam.

De Leeuw, E.D. (2008). Choosing the Method of Data Collection. In *International Handbook of Survey Methodology*, (Eds., E.D. De Leeuw, J.J. Hox and D.A. Dillman), Larence Erlbaum Associates, New York, U.S.A., 113-135.

De Leeuw, E., Callegaro, M., Hox, J., Korendijk, E. and Lensvelt-Mulders, G. (2007). The influence of advance letters on response in telephone surveys. *Public Opinion Quarterly*, 71, 413-443.

Dillman, D.A. (2007). Mail and internet surveys: The tailored design method, Wiley, Hoboken, New Jersey, U.S.A.

Greenberg, B.S., and Stokes, S.L. (1990). Developing an optimal call scheduling strategy for a telephone survey. *Journal of Official Statistics*, 6, 421-435.

Groves, R.M. (1989). Survey errors and survey costs, Wiley, New York, U.S.A.

Groves, R.M., and Couper, M.P. (1998). Nonresponse in Household Interview Surveys, Wiley series in probability and statistics, Survey methodology section.

Groves, R.M., Dillman, D., Eltinge, J. and Little, R. (2002). Survey Nonresponse, New York: Wiley Series in Probability and Statistics.

Groves, R.M., and Heeringa, S.G. (2006). Responsive design for household surveys: Tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society*, Series A, 169, 439-457.

Hartley, H.O., and Monroe, H. (1979). Interviewer assignments which minimize the effects of nonsampling errors. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 216-220.

Heyd, J.M., and Carlin, B.P. (1999). Adaptive design improvements in the continual reassessment method for Phase I studies. *Statistics in Medicine*, 18, 1307-1321.

Kalsbeek, W.D., Botman, S.L., Massey, J.T. and Liu, P.W. (1994). Cost-efficiency and the number of allowable call attempts in the National Health Interview Survey. *Journal of Official Statistics*, 10, 133-152.

Kersten, H.M.P., and Bethlehem, J.G. (1984). Exploring and reducing the non-response bias by asking the basic question, BPA 627-84-M1, CBS, Voorburg.

Kulka, R.A., and Weeks, M.F. (1988). Towards the development of optimal calling protocols for telephone surveys: A conditional probabilities approach. *Journal of Official Statistics*, 4, 319-322.

Laflamme, F., and Karaganis, M. (2010). Implementation of responsive collection design for CATI surveys at Statistics Canada, Paper presented at Q2010, 3-6 May, Helsinki, Finland.

Luiten, A., and Schouten, B. (2013). Adaptive fieldwork design to increase representative household survey response. *Journal of the Royal Statistical Society*, Series A, 176, 1, 169-190.

Luiten, A., and Wetzels, W. (2009). Indicators and data collection control. Work plan and preliminary findings pilot Statistics Netherlands, RISQ deliverable 8.2, www.risq-project.eu.

Lyberg, L.E., Biemer, P., Collins, M., de Leeuw, E.D., Dippo, C., Schwarz, N. and Trewin, D. (1997). Survey measurement and process quality, Wiley Series in Probability and Statistics, Wiley, Hoboken, New Jersey, U.S.A.

Lynn, P. (2003). PEDAKSI: Methodology for Collecting Data about Survey Non-Respondents. *Quality and Quantity*, Vol. 37, No. 3, 239-261.

Mohl, C., and Laflamme, F. (2007). Research and responsive design options for survey data collection at Statistics Canada. Proceedings of ASA Joint Statistical Meeting, Section 293, July 29 - August 2, Salt Lake City, U.S.A.

Moore, J. (1988). Self/Proxy response status and survey response quality. A review of literature. *Journal of Official Statistics*, 4, 2, 155-172.

Murphy, S.A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society*, Series B, 65, 331-355.

Peytchev, A., Riley, S., Rosen, J., Murphy, J. and Lindblad, M. (2010). Reduction of Nonresponse Bias in Surveys through Case Prioritization, Survey Research Methods, 4, 1, 21-29.

Phillips, O., and Tabuchi, T. (2009). Responsive design for the survey of labour and income dynamics. *Proceedings: Symposium 2009, Longitudinal Surveys: from design to analysis*, Statistics Canada.

Särndal, C.-E. (2011a). Three factors to signal non-response bias with applications to categorical auxiliary variables. *International Statistical Review*, 79, 2, 233-254.

Särndal, C.-E. (2011b). The 2010 Morris Hansen Lecture: Dealing with survey nonresponse in data collection, in estimation. *Journal of Official Statistics*, 27, 1, 1-21.

Särndal, C.-E., and Lundström, S. (2010). Design for estimation: Identifying auxiliary vectors to reduce nonresponse bias. *Survey Methodology*, 36, 2, 131-144.

Schouten, B., Cobben, F. and Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, 35, 1, 101-113.

Schouten, B., Luiten, A., Loosveldt, G., Beullens, K. and Kleven, Ø. (2010). Monitoring and changing data collection through R-indicators and partial R-indicators, RISQ Working paper. Available at www.risq-project.eu.

Schouten, J.G., Shlomo, N. and Skinner, C. (2011). Indicators for monitoring and improving representativeness of response. *Journal of Official Statistics*, 27, 2, 231-253.

Wagner, J. (2008). Adaptive survey design to reduce nonresponse bias, Ph.D. thesis, University of Michigan, U.S.A.

Zajonc, T. (2012). Bayesian inference for dynamic treatment regimes: Mobility, equity and efficiency in student tracking. *Journal of the American Statistical Association*, 107, 80-92.

Survey Methodology, June 2013
Vol. 39, No. 1, pp. 59-89
Statistics Canada, Catalogue No. 12-001-X

Automatic editing with hard and soft edits

Sander Scholtus[1]

Abstract

A considerable limitation of current methods for automatic data editing is that they treat all edits as hard constraints. That is to say, an edit failure is always attributed to an error in the data. In manual editing, however, subject-matter specialists also make extensive use of soft edits, *i.e.*, constraints that identify (combinations of) values that are suspicious but not necessarily incorrect. The inability of automatic editing methods to handle soft edits partly explains why in practice many differences are found between manually edited and automatically edited data. The object of this article is to present a new formulation of the error localisation problem which can distinguish between hard and soft edits. Moreover, it is shown how this problem may be solved by an extension of the error localisation algorithm of De Waal and Quere (2003).

Key Words: Automatic error localisation; Fellegi-Holt paradigm; Branch-and-bound algorithm; Numerical data; Categorical data; Mixed data.

1 Introduction

An important part of every statistical process is *data editing*, *i.e.*, detecting and correcting errors as well as missing values in the collected data. National statistical institutes have traditionally relied on manual editing, where the data is checked and, if necessary, adjusted by subject-matter experts. Unfortunately, this approach can be very time-consuming and expensive. Alternative methods have therefore been developed to increase the efficiency of the editing process, such as *selective editing* and *automatic editing*. This article focuses on the latter approach. We refer to De Waal, Pannekoek and Scholtus (2011) and their references for a discussion of selective editing and other forms of statistical data editing.

The goal of automatic editing is to accurately detect and correct errors as well as missing values in a data file in a fully automated manner, *i.e.*, without human intervention. Provided that automatic editing leads to data of sufficient quality, it can be used as a partial alternative to manual editing. In practice, automatic editing implies that the data is made consistent with respect to a set of constraints: the so-called *edits*. Examples of edits include:

1. Sander Scholtus, Statistics Netherlands, P.O. Box 24500, 2490 HA The Hague, The Netherlands. E-mail: sshs@cbs.nl.

$$\text{Profit} = \text{Total Turnover} - \text{Total Costs}; \tag{1.1}$$

and

$$\text{Profit} \leq 0.6 \times \text{Total Turnover}. \tag{1.2}$$

Most automatic editing methods proceed by solving two separate problems: first the *error localisation problem*, *i.e.*, determining which variables are erroneous or missing, and subsequently the *consistent imputation problem*, *i.e.*, determining new values for these variables that satisfy all edits. The present article focuses on the error localisation problem.

With respect to the two examples of edits given above, it is interesting to note the conceptual difference that exists between them. Edit (1.1) is an example of an edit that has to hold by definition, so that every combination of values that fails this edit necessarily contains an error. Edits of this type are known as *hard edits, fatal edits,* or *logical edits*. Edit (1.2), on the other hand, is an example of an edit that identifies combinations of values that are implausible but not necessarily incorrect. In this example, records for which *Profit* is larger than 60% of *Total Turnover* are considered suspicious. However, it is conceivable that such a combination of values is occasionally correct. Edits of this type, which do not identify errors with certainty, are known as *soft edits* or *query edits*.

An important limitation of existing algorithms for automatic editing is that they treat all edits as hard edits. That is to say, a failed edit is always attributed to an error in the data. In manual editing, however, subject-matter specialists also make extensive use of soft edits. During automatic editing, these soft edits are either not used at all, or else interpreted as hard edits. Both solutions are unsatisfactory: in the first case some errors may be missed during automatic editing, and in the second case some correct values may be wrongfully identified as erroneous. In fact, the inability of automatic editing methods to handle soft edits partly explains why in practice many differences are found between manually edited and automatically edited data.

The object of this article is to present a new formulation of the automatic error localisation problem which can distinguish between hard edits and soft edits. In addition, the article shows

how the error localisation algorithm of SLICE – the software package for automatic editing developed at Statistics Netherlands – can be adapted to solve this new error localisation problem.

The remainder of this article is organised as follows. Section 2 provides a brief summary of existing methods for solving the error localisation problem. A distinction between hard and soft edits is introduced in the error localisation problem in Section 3. Section 4 extends the theory behind the algorithm of SLICE to the case that not all edits have to be satisfied. Based on these theoretical results, an algorithm that solves the error localisation problem for hard and soft edits is outlined in Section 5. In Section 6, the new algorithm is illustrated by means of a small example. Section 7 mentions the first experiences with a practical implementation of the new algorithm. Finally, some concluding remarks follow in Section 8.

2 Background

2.1 Edits

The problem to be discussed in this article entails, in its most general form, the detection of erroneous and missing values in a record containing both categorical variables $(v_1,\ldots,v_m)$ and numerical variables $(x_1,\ldots,x_p)$. These variables are supposed to satisfy a set of restrictions (edits), each of which can be written in one of the following forms:

$$\begin{aligned}\psi^k\text{: IF } &(v_1,\ldots,v_m) \in F_1^k \times \cdots \times F_m^k \\ &\text{THEN } (x_1,\ldots,x_p) \in \{a_{k1}x_1 + \cdots + a_{kp}x_p + b_k \geq 0\}\end{aligned} \tag{2.1}$$

or

$$\begin{aligned}\psi^k\text{: IF } &(v_1,\ldots,v_m) \in F_1^k \times \cdots \times F_m^k \\ &\text{THEN } (x_1,\ldots,x_p) \in \{a_{k1}x_1 + \cdots + a_{kp}x_p + b_k = 0\}.\end{aligned} \tag{2.2}$$

In these expressions, F_j^k is a subset of D_j, the domain of observed values for the categorical variable v_j, and a_{kj} and b_k are known numerical constants. The index k is used to number the edits. Note that D_j is assumed to contain all values of v_j that may be encountered in practice; this includes erroneous values. To simplify matters, we restrict the problem to edits having linear

numerical conditions. This class of edits turns out to be sufficiently powerful for most practical applications (*cf*. De Waal 2005).

A record $(v_1^0, \ldots, v_m^0, x_1^0, \ldots, x_p^0)$ is said to *fail* an edit if the categorical IF-condition is true (*i.e.*, $v_j^0 \in F_j^k$ for all $j = 1, \ldots, m$), but the numerical THEN-condition is false (*i.e.*, either $a_{k1}x_1 + \cdots + a_{kp}x_p + b_k < 0$ or $a_{k1}x_1 + \cdots + a_{kp}x_p + b_k \neq 0$, depending on the form of the edit). Otherwise, we say that the edit is *satisfied* by that record. It should be noted that an edit is always satisfied by any record for which the categorical IF-condition is false, regardless of the status of the numerical THEN-condition. A record is called *consistent* if it satisfies every edit.

A categorical variable v_j is said to be *involved* in an edit if and only if $F_j^k \neq D_j$, since any edit with $F_j^k = D_j$ is failed or satisfied regardless of the value of v_j. Similarly, a numerical variable x_j is said to be involved in an edit if and only if $a_{kj} \neq 0$. We may assume that $F_j^k \neq \varnothing$, where $\varnothing$ denotes the empty set. Clearly, a degenerate edit with $F_j^k = \varnothing$ can be discarded with no loss of information, since it is never failed. The same holds for any edit with a numerical THEN-condition that is always true.

Two important special cases of (2.1) and (2.2) are edits that involve only categorical or only numerical variables. A purely categorical edit has the following form:

$$\psi^k\text{: IF } (v_1, \ldots, v_m) \in F_1^k \times \cdots \times F_m^k \text{ THEN } \varnothing. \tag{2.3}$$

Edit (2.3) is failed by each record for which the categorical condition is true. A purely numerical edit can be written as:

$$\psi^k\text{: } (x_1, \ldots, x_p) \in \{a_{k1}x_1 + \cdots + a_{kp}x_p + b_k \geq 0\} \tag{2.4}$$

or

$$\psi^k\text{: } (x_1, \ldots, x_p) \in \{a_{k1}x_1 + \cdots + a_{kp}x_p + b_k = 0\}. \tag{2.5}$$

Edits (2.4) and (2.5) are failed by each record for which the numerical conditions are false.

Edits (1.1) and (1.2) above are examples of purely numerical edits. A simple example of a purely categorical edit is:

$$\text{IF (Age, Marital Status)} \in \{\text{“<16”}\} \times \{\text{“Married”}\} \text{ THEN } \varnothing.$$

This edit states that persons aged less than 16 years cannot be married. Finally, an example of a mixed edit is:

$$\text{IF Age} \in \{\text{“<12”}\} \text{ THEN Income} = 0.$$

According to this edit, persons aged less than 12 years do not have a positive income.

2.2 The error localisation problem

For a given record $(v_1^0, \ldots, v_m^0, x_1^0, \ldots, x_p^0)$ and a collection of edits, it is straightforward to verify which values in the record are missing and whether any of the edits are failed. However, given that some of the edits are failed, solving the error localisation problem is much less straightforward. On the one hand, most edits involve more than one variable, and on the other hand, most variables are involved in more than one edit.

In order to solve the error localisation problem automatically, one has to adopt a formal strategy for finding erroneous values. The most commonly-used strategy is based on the paradigm of Fellegi and Holt (1976): make the record consistent by changing the smallest possible number of original values. Other strategies have also been proposed; for instance, Little and Smith (1987) suggested a criterion based on outlier detection (without edits) and Casado Valero, Del Castillo Cuervo-Arango, Mateo Ayerra and De Santos Ballesteros (1996) formulated error localisation as a quadratic minimisation problem. We shall restrict attention to the Fellegi-Holt paradigm here, because of its frequent use in official statistics.

The original Fellegi-Holt paradigm is easily generalised to allow a distinction between a priori suspicious and less suspicious variables. To this end, one associates a *confidence weight* to each variable. According to the generalised Fellegi-Holt paradigm, one should search for a subset of

the variables which (*i*) can be imputed in such a way that the imputed record satisfies all edits, and (*ii*) minimises the following target function:

$$D_{\mathrm{FH}} = \sum_{j=1}^{m} w_j^C y_j^C + \sum_{j=1}^{p} w_j^N y_j^N . \tag{2.6}$$

Here, w_j^C and w_j^N denote the confidence weights of the categorical and numerical variables, respectively. The binary target variables y_j^C and y_j^N describe the structure of the solution: $y_j^C = 1$ if v_j is to be imputed and $y_j^C = 0$ otherwise, and similarly $y_j^N = 1$ if x_j is to be imputed and $y_j^N = 0$ otherwise. Since variables with missing values have to be imputed with certainty, we set $y_j^C = 1$ or $y_j^N = 1$ when v_j^0 or x_j^0 is missing.

Fellegi and Holt (1976) also presented a method for solving the error localisation problem under this paradigm. This method first derives a well-defined set of logically implied edits from the original set of edits, to obtain a so-called *complete set of edits*. Next, the error localisation problem may be formulated as a straightforward set-covering problem for any record (Fellegi and Holt 1976; Boskovitz, Goré and Wong 2005). Unfortunately, especially for numerical data the complete set of edits can be extremely large in practice, so the method of Fellegi and Holt is not always computationally feasible.

Many alternative algorithms have been developed for error localisation according to the Fellegi-Holt paradigm. Besides improvements on Fellegi and Holt's original method (Garfinkel, Kunnathur and Liepins 1986; Winkler 1995), the list includes formulations based on vertex generation (Sande 1978; Kovar and Whitridge 1990; Todaro 1999; De Waal 2003a), cutting planes (Garfinkel *et al.* 1986; Garfinkel, Kunnathur and Liepins 1988; Ragsdale and McKeown 1996), and mixed integer (Schaffer 1987; Riera-Ledesma and Salazar-González 2003) and integer programming (Bruni 2004 and 2005); see also De Waal *et al.* (2011) for an overview. Here, we shall focus on a branch-and-bound algorithm due to De Waal and Quere (2003) which, in contrast to some of the above approaches, can handle a mix of categorical and numerical data.

This algorithm has been implemented in the software package SLICE at Statistics Netherlands and has been found to be computationally feasible in practice.

2.3 The branch-and-bound algorithm of SLICE

A detailed description of the error localisation algorithm implemented in SLICE can be found in De Waal and Quere (2003), De Waal (2003b), and De Waal *et al.* (2011). Here, we only mention those aspects of the algorithm that we shall need later. For a general introduction to branch-and-bound algorithms, see *e.g.*, Nemhauser and Wolsey (1988).

For each record, the SLICE algorithm sets up a binary tree, as illustrated in Figure 2.1. In the root node of the tree, we start with the original set of edits and we select one of the variables. From the root node, two branches are added to the tree. In the first branch, the original value of the selected variable in the record is assumed to be correct, and in the second branch this value is assumed to be erroneous. Both assumptions correspond with a transformation of the set of edits, to be outlined below, after which the selected variable is no longer involved in the edits: the selected variable has been *treated*. Next, one of the remaining variables is selected and the operation is repeated.

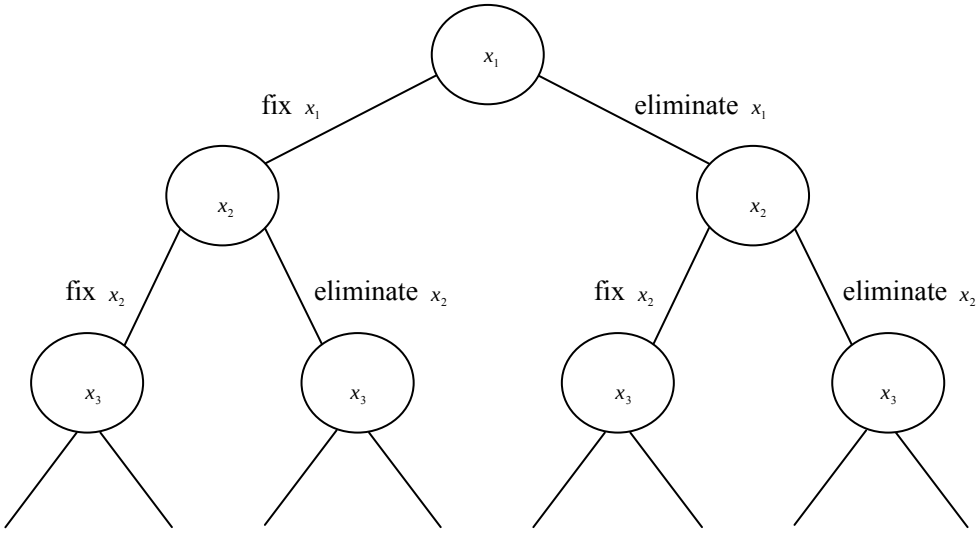


Figure 2.1 Illustration of the branch-and-bound algorithm as a binary tree

Once all variables have been treated, the algorithm reaches an end node of the tree. It is seen that, together, the end nodes of the binary tree enumerate all possible choices of erroneous subsets of variables. The transformed set of edits corresponding to an end node does not involve any variables, so it must either be empty or consist of elementary relations such as "$1 \geq 0$" (a tautology) and "$-1 \geq 0$" (a self-contradicting statement). As will be discussed below, it is possible to satisfy the original edits by only imputing the variables that have been considered erroneous in the branch leading to an end node, if and only if the transformed set of edits for that end node contains no self-contradicting statements. Using this property, all feasible solutions to the error localisation problem may be identified. Moreover, since we are only interested in feasible solutions that minimise target function (2.6), a branch of the tree may be pruned as soon as we find that it only leads to end nodes corresponding with infeasible or suboptimal solutions.

We will now outline the transformations of the set of edits that occur, depending on whether a variable is assumed to be correct or erroneous. A variable that is assumed to be correct is removed from the edits by simply substituting the original value from the record in the edits. This is called *fixing* a variable to its original value. A variable that is assumed to be erroneous is removed from the edits by a more complex operation, called *eliminating* a variable from the edits. Numerical variables and categorical variables are eliminated by two different but equivalent methods.

To eliminate a numerical variable, say x_g, from a set of edits having the general forms (2.1) and (2.2), we generate logically implied edits by considering all pairs of edits ψ^s and ψ^t that involve x_g. We first check whether $F_j^s \cap F_j^t \neq \varnothing$ for all $j = 1, \ldots, m$; if any of these intersections yields the empty set, then the pair ψ^s and ψ^t does not generate an implied edit. If the numerical THEN-condition of one of the edits, say ψ^s, is an equality, then this equality may be solved for x_g. By substituting the resulting expression for x_g in the THEN-condition of ψ^t, we obtain the numerical THEN-condition of the implied edit. The categorical IF-condition of the implied edit is found by taking the non-empty intersections $F_j^* = F_j^s \cap F_j^t$ for $j = 1, \ldots, m$.

If the numerical THEN-conditions of ψ^s and ψ^t are both inequalities, the algorithm uses a technique called *Fourier-Motzkin elimination* (see *e.g.*, Williams 1986) to generate an implied edit. A pair of edits is relevant for this elimination method only if the coefficients of x_g have opposite signs, so we may assume without loss of generality that $a_{sg} < 0$ and $a_{tg} > 0$. The implied edit generated from ψ^s and ψ^t may then be written as (*cf*. De Waal and Quere 2003):

$$\begin{aligned} \psi^*\text{: IF } & (v_1, \ldots, v_m) \in F_1^* \times \cdots \times F_m^* \\ & \text{THEN } (x_1, \ldots, x_p) \in \{a_1^* x_1 + \cdots + a_p^* x_p + b^* \geq 0\} \end{aligned} \tag{2.7}$$

with $a_j^* = a_{tg}a_{sj} - a_{sg}a_{tj}, b^* = a_{tg}b_s - a_{sg}b_t,$ and $F_j^* = F_j^s \cap F_j^t$ as above. This edit does not involve x_g, since $a_g^* = 0$. In this manner, implied edits are generated by considering all pairs of edits that involve x_g. These edits are added to the set of original edits that do not involve x_g, to find the transformed set of edits obtained by eliminating x_g.

For the elimination of categorical variables, De Waal and Quere (2003) make the simplifying assumption that these variables are only selected when all numerical variables have been treated. This assumption implies that categorical variables are always eliminated from purely categorical edits of the form (2.3). To eliminate a categorical variable, say v_g, from a set of edits of the form (2.3), a technique is used that was first described by Fellegi and Holt (1976).

Consider all minimal sets of edits T with the following properties:

$$F_g^*(T) = \bigcup_{k \in T} F_g^k = D_g \tag{2.8}$$

and

$$F_j^*(T) = \bigcap_{k \in T} F_j^k \neq \varnothing, \text{ for } j = 1, \ldots, g-1, g+1, \ldots, m. \tag{2.9}$$

Here, by "minimal" we mean that property (2.8) does not hold for any set $T' \subset T$. Each of these minimal sets T generates an implied edit:

$$\psi^*\text{: IF } (v_1, \ldots, v_m) \in F_1^*(T) \times \cdots \times F_m^*(T) \text{ THEN } \varnothing, \tag{2.10}$$

which does not involve v_g because of property (2.8). These implied edits are added to the set of original edits that do not involve v_g, to find the transformed set of edits obtained by eliminating v_g.

It should be clear that the computational work of the algorithm lies mainly in the elimination steps. In particular, it is known that the number of implied edits under Fourier-Motzkin elimination may be exponential in the number of eliminated variables (Schrijver 1986).

A fundamental property of both elimination techniques, for numerical and categorical variables, is exhibited by the following result. Consider a system of implied edits Ψ_1 obtained by eliminating x_g or v_g from a system of edits Ψ_0. Then the original values of the untreated variables satisfy all edits in Ψ_1, if and only if there exists a value for x_g or v_g that, together with these original values, satisfies all edits in Ψ_0. For a proof, see Theorem 8.1 in De Waal (2003b) or Theorem 4.3 in De Waal *et al.* (2011). The above-mentioned correspondence between end nodes without self-contradicting elementary relations and feasible solutions to the error localisation problem follows from a repeated application of this fundamental property.

3 An error localisation problem with hard and soft edits

In the formulation of the error localisation problem given in Section 2.2, which is based on the Fellegi-Holt paradigm, it is tacitly assumed that all edits are hard edits. Consequently, the only subsets of the variables that are considered feasible solutions to this problem are those which can be imputed to make the record consistent with respect to all edits. As mentioned in the introduction, this interpretation of all edits as hard edits can lead to systematic differences between automatic editing and manual editing, because it precludes a meaningful use of soft edits. In this section, we suggest a new formulation of the error localisation problem which distinguishes between hard and soft edits.

Let Ψ denote the set of edits to be used in the error localisation problem. We assume that this set can be partitioned into two disjoint subsets: $\Psi = \Psi_H \cup \Psi_S$. The edits in Ψ_H are hard edits;

the edits in Ψ_S are soft edits. From now on, a subset of the variables is considered a feasible solution to the error localisation problem if it can be imputed to satisfy all edits in Ψ_H. Moreover, we want to use the status of the imputed record with respect to the edits in Ψ_S as auxiliary information in the choice of an optimal solution. This may be done by adding another term to (2.6).

More precisely, the objective of the new error localisation problem is to find a subset of the variables which (*i*) can be imputed so that the imputed record satisfies all edits in Ψ_H, and (*ii*) minimises the following target function:

$$D = \lambda D_{\text{FH}} + (1 - \lambda) D_{\text{soft}}, \tag{3.1}$$

where D_{soft} represents the costs associated with failed edits in Ψ_S. The parameter $\lambda \in [0,1]$ determines the relative contribution of both terms in (3.1). The original Fellegi-Holt paradigm is recovered as a special case by choosing $\lambda = 1$. Thus, the new error localisation problem can be seen as a generalisation of the old one.

In order to use (3.1) in practice, one has to choose an expression for D_{soft}. Probably the easiest way to assign costs to failed soft edits is to associate a fixed *failure weight* s_k to each edit in Ψ_S, and to define D_{soft} as the sum of the failure weights of the soft edits that remain failed:

$$D_{\text{soft}} = \sum_{k=1}^{K_S} s_k z_k, \tag{3.2}$$

with K_S the number of edits in Ψ_S and z_k a binary variable such that $z_k = 1$ if the k^{th} soft edit is failed and $z_k = 0$ otherwise. The failure weights may be chosen by subject-matter experts, analogously to the confidence weights, to express the importance that is attached to different soft edits from a subject-matter related point of view. Alternatively, the failure weights may be based on the proportion of records that fail each soft edit in a historical data set which has been edited manually.

A drawback of using fixed failure weights is that they do not take the size of the edit failures into account: every record that fails a particular soft edit receives the same contribution to D_{soft},

namely s_k. By contrast, a human editor sees a soft edit failure as an indication that an observed combination of values is suspicious, and the degree of suspicion depends on the size of the edit failure: a small failure is ignored more easily than a large failure. Hence, it seems interesting to take the size of the edit failures into account in D_{soft}. This point will be taken up in Section 8, since it introduces certain additional difficulties. For now, we assume that expression (3.2) is used.

We should mention that taking soft restrictions into account by adding an appropriate term to a target function is a well-known technique in mathematical optimisation. The idea is related to other optimisation techniques, such as Lagrangian relaxation (see *e.g.*, Nemhauser and Wolsey 1988). One example of a practical application with soft constraints is that of the so-called benchmarking problem for national accounts (Magnus, Van Tongeren and De Vos 2000). To our best knowledge, the application in the context of the error localisation problem is new.

We should also note that expression (3.1) is in some respects similar to the minimisation criterion of the Nearest-neighbour Imputation Methodology (NIM) developed by Statistics Canada for editing demographic census data (Bankier, Lachance and Poirier 2000; Bankier and Crowe 2009). In particular, the NIM also departs from the Fellegi-Holt paradigm by minimising a convex combination of two terms, the first measuring the amount of imputation and the second measuring the plausibility of the imputed record.

4 A short theory of edit failures

4.1 Numerical data

Having formulated a new error localisation problem, we will now show how this problem may be solved by an adapted version of the branch-and-bound algorithm of De Waal and Quere (2003). To do this, we first need to extend the fundamental property mentioned at the end of Section 2.3 to the case that some of the edits may be failed. For convenience, we shall first

examine the case of purely numerical data. The next subsection examines the case of purely categorical and mixed data.

In the case of purely numerical data, all edits take the form (2.4) or (2.5). Moreover, the implied edit (2.7) is reduced to its numerical part. The fundamental property given at the end of Section 2.3 implies in particular the following: if a given set of values for $x_1, \ldots, x_{g-1}, x_{g+1}, \ldots, x_p$ does not satisfy the implied edit (2.7), then it is impossible to find a value for x_g that satisfies ψ^s and ψ^t simultaneously. However, it is still possible in this case to find a value for x_g that satisfies one of the edits ψ^s or ψ^t. This observation, which is more or less trivial, forms the basis for the proof of Theorem 1 below.

Suppose that, at some point during an execution of the branch-and-bound algorithm of De Waal and Quere (2003), q numerical variables have been treated (*i.e.*, either eliminated or fixed). We denote the current set of edits by Ψ_q, and the edits in this set by ψ_q^k. By definition, $\Psi_0 \equiv \Psi$, the original set of edits. It is possible to associate with each current edit ψ_q^k an index set B_q^k, which contains the indices of all the original edits that have been used, directly or indirectly, to derive this edit. In fact, B_q^k can be defined recursively as follows:

- For an original edit ψ_0^k, we define $B_0^k := \{k\}$.
- For an edit ψ_q^k which is derived from one other edit ψ_{q-1}^l, either by fixing a variable to its original value or by simply copying the edit, we define $B_q^k := B_{q-1}^l$.
- For an edit ψ_q^k which is derived by eliminating a variable from a set of edits ψ_{q-1}^t $(t \in T)$, we define $B_q^k := \bigcup_{t \in T} B_{q-1}^t$.

Note that, for numerical data, the set T in the last item always contains exactly two edits. Larger edit sets may be encountered in the categorical case considered below.

A set B is called a *representing set* of a collection of sets $B_q^{k_1}, \ldots, B_q^{k_r}$ if it contains at least one element from each of $B_q^{k_1}, \ldots, B_q^{k_r}$; see, for instance, Mirsky (1971, page 25). It should be noted that, in our case, a representing set B identifies a subset of Ψ_0, the set of original edits. We can now formulate the following theorem.

Theorem 1. Suppose that q numerical variables have been treated and that the current set of numerical edits can be partitioned as $\Psi_q = \Psi_q^{(1)} \cup \Psi_q^{(2)}$, where the edits in $\Psi_q^{(1)}$ are satisfied by the original values of the $p - q$ remaining variables, and the edits in $\Psi_q^{(2)}$ are failed. Let B be a representing set of the index sets B_q^k for all $\psi_q^k \in \Psi_q^{(2)}$. Then there exist values for the eliminated variables that, together with the original values of the other variables, satisfy all original edits except those in B.

Proof. The proof of this theorem is given in Appendix A.1.

Example: Suppose that there are three numerical variables (x_1, x_2, x_3) that should satisfy the following eight edits:

$$\begin{aligned}
\psi_0^1&: \quad x_1 + x_2 + x_3 = 20\\
\psi_0^2&: \quad x_1 - x_2 \geq 3\\
\psi_0^3&: \quad -x_1 + x_2 \geq -6\\
\psi_0^4&: \quad -x_1 + x_3 \geq 5\\
\psi_0^5&: \quad x_1 - x_3 \geq -10\\
\psi_0^6&: \quad x_1 \geq 0\\
\psi_0^7&: \quad x_2 \geq 0\\
\psi_0^8&: \quad x_3 \geq 0.
\end{aligned}$$

The record $(x_1^0, x_2^0, x_3^0) = (10, 1, -3)$ is inconsistent with respect to these edits. Upon eliminating x_1 from the original set of edits, we find the following updated set of edits:

$$\begin{aligned}
\psi_1^1&: \quad -2x_2 - x_3 \geq -17 \quad (B_1^1 = \{1,2\})\\
\psi_1^2&: \quad 2x_2 + x_3 \geq 14 \quad (B_1^2 = \{1,3\})\\
\psi_1^3&: \quad x_2 + 2x_3 \geq 25 \quad (B_1^3 = \{1,4\})\\
\psi_1^4&: \quad -x_2 - 2x_3 \geq -30 \quad (B_1^4 = \{1,5\})\\
\psi_1^5&: \quad -x_2 - x_3 \geq -20 \quad (B_1^5 = \{1,6\})\\
\psi_1^6&: \quad x_2 \geq 0 \quad (B_1^6 = \{7\})\\
\psi_1^7&: \quad x_3 \geq 0 \quad (B_1^7 = \{8\})\\
\psi_1^8&: \quad 0 \geq -3 \quad (B_1^8 = \{2,3\})\\
\psi_1^9&: \quad -x_2 + x_3 \geq 8 \quad (B_1^9 = \{2,4\})\\
\psi_1^{10}&: \quad x_2 - x_3 \geq -16 \quad (B_1^{10} = \{3,5\})\\
\psi_1^{11}&: \quad 0 \geq -5 \quad (B_1^{11} = \{4,5\})\\
\psi_1^{12}&: \quad x_2 \geq -6 \quad (B_1^{12} = \{3,6\})\\
\psi_1^{13}&: \quad x_3 \geq 5 \quad (B_1^{13} = \{4,6\}).
\end{aligned}$$

The index set B_1^k is displayed in brackets next to each edit.

By substituting the original values of x_2 and x_3 in the current set of edits, we see that $\psi_1^2, \psi_1^3, \psi_1^7, \psi_1^9,$ and ψ_1^{13} are failed. The set $B = \{1, 4, 8\}$ is a representing set for the associated index sets B_1^k. According to Theorem 1, there exists a value for x_1 which, together with the original values of x_2 and x_3, satisfies the original edits apart from $\psi_0^1, \psi_0^4,$ and ψ_0^8. That this assertion is correct can be seen by substituting $x_2^0 = 1$ and $x_3^0 = -3$ into the original set of edits; in fact, any value $x_1 \in [4, 7]$ will do.

The importance of Theorem 1 is that it enables one to evaluate, at each node of the branch-and-bound algorithm, which combinations of the original edits could be satisfied by imputing the variables that have been eliminated so far, and also which edits would remain failed. In particular, if we distinguish between hard and soft original edits, then this result makes it possible to use the branch-and-bound algorithm to find all feasible solutions to the new error localisation problem from Section 3, and also to evaluate, for each feasible solution, which of the soft edits remain failed, and hence to evaluate the value of D_{soft}. This idea will be elaborated in Section 5.

4.2 Categorical and mixed data

We shall now derive a similar result to Theorem 1 for the case of purely categorical data. At the end of this section, we shall combine the two results so that they may also be applied to mixed data.

In the case of purely categorical data, all edits take the form (2.3). Let us consider the elimination method for categorical variables described in Section 2.3. If a given set of values for $v_1, \ldots, v_{g-1}, v_{g+1}, \ldots, v_m$ does not satisfy the implied edit (2.10), then it is not possible to find a value for v_g that, together with the other values, satisfy all edits ψ^k with $k \in T$ simultaneously. This is true because, by property (2.9), $F_j^*(T) \subseteq F_j^k$ for all $j \neq g$ and all $k \in T$. Hence, if (2.10) is failed by $v_1, \ldots, v_{g-1}, v_{g+1}, \ldots, v_m,$ then plugging these values into an original edit with

$k \in T$ produces a non-degenerate univariate edit for v_g. Moreover, every possible value of v_g fails at least one of these univariate edits, because of property (2.8). Interestingly, it is still always possible in this case to find a value for v_g that satisfies all edits in T but one. This follows from property (2.8) and the fact that T is a minimal set having this property: for each $k \in T, F_g^k$ must contain at least one value from D_g that is not covered by any other F_g^l with $l \in T$.

We now present the analogue of Theorem 1 for categorical data, using the same notation as for numerical data. In particular, the recursive definition of B_q^k is exactly the same as in Section 4.1.

Theorem 2. Suppose that q categorical variables have been treated and that the current set of categorical edits can be partitioned as $\Psi_q = \Psi_q^{(1)} \cup \Psi_q^{(2)}$, where the edits in $\Psi_q^{(1)}$ are satisfied by the original values of the $m - q$ remaining variables, and the edits in $\Psi_q^{(2)}$ are failed. Let B be a representing set of the index sets B_q^k for all $\psi_q^k \in \Psi_q^{(2)}$. Then there exist values for the eliminated variables that, together with the original values of the other variables, satisfy all original edits except those in B.

Proof. The proof of this theorem is given in Appendix A.2.

For an example that illustrates the use of this theorem, see Scholtus (2011).

Finally, we remark that Theorem 1 and Theorem 2 can be used together when the data is a mix of categorical and numerical variables. This follows from the structure of the branch-and-bound algorithm of De Waal and Quere (2003), where categorical variables are only treated once all numerical variables have been eliminated or fixed. Hence, the two results may be applied consecutively. There is a slight difference in the procedure for eliminating numerical variables, namely that implied edits are only generated from pairs of edits having an overlapping IF-condition; see Section 2.3. However, this does not affect the correctness of Theorem 1.

5 Solving the error localisation problem with hard and soft edits

We shall now describe an adapted version of the branch-and-bound algorithm of De Waal and Quere, which may be used to solve the error localisation problem defined in Section 3. The basic setup of the algorithm is the same as in Section 2.3. In particular, the procedures for eliminating and fixing variables are carried out the same way as in the original algorithm.

The main difference is that now in each node, the current set of edits Ψ_q is partitioned into a current set of hard edits Ψ_{qH} and a current set of soft edits Ψ_{qS}. For the root node, the partition simply follows that of the original set of edits, *i.e.*, $\Psi_{0H} = \Psi_H$ and $\Psi_{0S} = \Psi_S$. For all other nodes, the partition can be summarised as follows: if an edit is generated only from hard edits, then it is a hard edit; if any soft edits are involved in its generation, then it is a soft edit. Furthermore, for each soft edit $\psi_{qS}^k \in \Psi_{qS}$, we construct an index set B_{qS}^k – analogous to B_q^k in Section 4 – which contains the indices of all the original *soft* edits ψ_{0S}^k that were involved, directly or indirectly, in its generation.

Having generated Ψ_{qH} and Ψ_{qS} for a particular node, we can fill in the original values of the variables that have not been treated yet, to check which of these edits are failed. In the old algorithm, this check could have two possible outcomes: either more variables need to be eliminated (at least one of the edits is failed), or a feasible solution has been found (none of the edits are failed). In the new algorithm, three different situations may arise.

First of all, if at least one edit in Ψ_{qH} is failed, then the variables that have been eliminated so far cannot be imputed to satisfy the original hard edits. Hence, more variables need to be eliminated. In this case, we continue the generation of branches from the current node.

A second possibility is that none of the edits in Ψ_{qH} or Ψ_{qS} are failed. This means that the variables that have been eliminated so far can be imputed to satisfy all the original edits, both hard and soft. Thus a feasible solution has been found, for which the value of target function (3.1) equals $D = \lambda D_{\mathrm{FH}}$. If this value is smaller than or equal to the value of (3.1) for the best solution found so far, say $D_{\min}$, then the new solution is stored. Otherwise, it is discarded. Either way, it

is not useful to continue the algorithm from the current node, because if more variables are eliminated, the value of D can only increase. Hence, we return to the last previous branch that has not been completely searched yet and continue the algorithm from there.

The last possibility is that the edits in Ψ_{qH} are satisfied, but that at least one edit in Ψ_{qS} is failed. In this case, the variables that have been eliminated so far can be imputed to satisfy the original hard edits, but not all the original soft edits. Hence, a feasible solution to the error localisation problem has been found, but the contribution of D_{soft} to D is non-zero. According to Theorem 1 or Theorem 2, it is possible to satisfy all original soft edits, except those in a representing set B of the index sets B_{qS}^{k} for all failed edits in Ψ_{qS}. Since this property is shared by all representing sets, we are free to choose B in such a way that D_{soft} is minimised, given the selection of variables to impute. If expression (3.2) is used for D_{soft}, then the optimal choice of B can be found by solving the following minimisation problem:

$$\begin{aligned} &\min \sum_{k=1}^{K_S} s_k z_k, \text{ under the conditions that:} \\ &\sum_{k \in B_{qS}^{l}} z_k \geq 1, \text{ for all failed } \psi_{qS}^{l} \in \Psi_{qS}, \\ &z_k \in \{0,1\}, k = 1, \ldots, K_S. \end{aligned} \tag{5.1}$$

This is a standard binary linear optimisation problem for which algorithms are available (see *e.g.*, Nemhauser and Wolsey 1988). The solution consists of a vector $(z_1^*, \ldots, z_{K_S}^*)$ of zeros and ones. The associated optimal representing set is $B^* = \{k: z_k^* = 1\}$ and the associated contribution of D_{soft} to D is precisely the minimal value of problem (5.1), say

$$D_{\text{soft}}^* = \sum_{k=1}^{K_S} s_k z_k^* = \sum_{k \in B^*} s_k.$$

As in the previous case, the value $D = \lambda D_{\text{FH}} + (1 - \lambda) D_{\text{soft}}^*$ is compared to $D_{\min}$. If $D \leq D_{\min}$, then the current solution is stored, otherwise it is discarded. Either way, it is meaningful in this case to continue the algorithm from the current node, because eliminating more variables may lead to a lower value of the target function. This can happen because a solution that imputes more

variables typically fails fewer soft edits. Therefore, we continue the generation of branches from the current node.

The correctness of this algorithm follows from the correctness of the original algorithm of De Waal and Quere (2003) and the theory presented in Section 4. The index sets B_q^k only have to be computed for the soft edits, because a subset of the variables is never considered a feasible solution to the error localisation problem when at least one of the hard edits remains failed. This means that, in every application of Theorem 1 or Theorem 2, all implied edits in Ψ_{qH} must be contained in $\Psi_q^{(1)}$. Finally, we note that the new algorithm reduces to the original algorithm of De Waal and Quere (2003) in the special case that no soft edits have been specified.

6 Example

To illustrate the algorithm of Section 5, we will apply it to a small example with numerical data. This is essentially an example from De Waal (2003b) to which we have added a distinction between hard and soft edits. For a somewhat larger example involving a mix of categorical and numerical variables, see Scholtus (2011).

In a fictitious business survey, there are four numerical variables: *total turnover* (T), *profit* (P), *total costs* (C), and *number of employees* (N). The following hard edits and soft edits have been identified:

$$\begin{aligned}
\psi_{0H}^1&: \quad T - C - P = 0 \\
\psi_{0H}^2&: \quad T \geq 0 \\
\psi_{0H}^3&: \quad C \geq 0 \\
\psi_{0H}^4&: \quad N \geq 0 \\
\psi_{0H}^5&: \quad 550N - T \geq 0 \\
\psi_{0S}^1&: \quad 0.5T - P \geq 0 \quad (B_{0S}^1 = \{1\}) \\
\psi_{0S}^2&: \quad P + 0.1T \geq 0 \quad (B_{0S}^2 = \{2\}).
\end{aligned}$$

Consider the following unedited record: $(T^0, P^0, C^0, N^0) = (100; 40{,}000; 60{,}000; 5)$. This record fails the first hard edit and the first soft edit. The confidence weights of the variables are

$(w_T, w_P, w_C, w_N) = (2, 1, 1, 3)$. We choose the failure weights of the two soft edits to be $s_1 = s_2 = 2$. Finally, we choose $\lambda = 1 / 2$ in expression (3.1).

Suppose that the variable P is selected first. In the branch where P is eliminated from the original edits, we obtain the following new set of edits:

$$\begin{array}{lrll}
\psi_{1H}^1: & T \geq 0 & & (\psi_{0H}^2) \\
\psi_{1H}^2: & C \geq 0 & & (\psi_{0H}^3) \\
\psi_{1H}^3: & N \geq 0 & & (\psi_{0H}^4) \\
\psi_{1H}^4: & 550N - T \geq 0 & & (\psi_{0H}^5) \\
\psi_{1S}^1: & -0.5T + C \geq 0 & (B_{1S}^1 = \{1\}) & (\psi_{0H}^1, \psi_{0S}^1) \\
\psi_{1S}^2: & 1.1T - C \geq 0 & (B_{1S}^2 = \{2\}) & (\psi_{0H}^1, \psi_{0S}^2) \\
\psi_{1S}^3: & 0.6T \geq 0 & (B_{1S}^3 = \{1, 2\}) & (\psi_{0S}^1, \psi_{0S}^2).
\end{array}$$

We have indicated in brackets from which of the previous edits each new edit is derived. The third soft edit ψ_{1S}^3 is in fact equivalent to the first hard edit ψ_{1H}^1, which means that it can be discarded.

Upon substituting the original values $(T^0, C^0, N^0) = (100; 60{,}000; 5)$ into the current edits, it is seen that all edits are satisfied except for ψ_{1S}^2. Since all hard edits are satisfied, identifying only the original value of P as erroneous is a feasible solution to the error localisation problem. Moreover, since $B = \{2\}$ is (trivially) a minimal representing set of B_{1S}^2, it is possible to impute a value for P which satisfies all the original edits except for ψ_{0S}^2. Hence, the value of target function (3.1) for this solution is $(w_P + s_2) / 2 = 3 / 2$.

Possibly, the current solution may be improved by eliminating another variable, say C, from the current set of edits. This yields:

$$\begin{array}{lrll}
\psi_{2H}^1: & T \geq 0 & & (\psi_{1H}^1) \\
\psi_{2H}^2: & N \geq 0 & & (\psi_{1H}^3) \\
\psi_{2H}^3: & 550N - T \geq 0 & & (\psi_{1H}^4) \\
\psi_{2S}^1: & 1.1T \geq 0 & (B_{2S}^1 = \{2\}) & (\psi_{1H}^2, \psi_{1S}^2) \\
\psi_{2S}^2: & 0.6T \geq 0 & (B_{2S}^2 = \{1, 2\}) & (\psi_{1S}^1, \psi_{1S}^2).
\end{array}$$

Each of the two new soft edits is redundant, because both are equivalent to hard edit ψ_{2H}^1. In fact, the remaining original values $(T^0, N^0) = (100, 5)$ satisfy all the current edits. This means

that P and C can be imputed to satisfy all the original edits, both hard and soft. The value of target function (3.1) for this solution equals $(w_P + w_C) / 2 = 1$. Thus, the new solution improves on the previous one. Moreover, this solution cannot be improved further by eliminating more variables in the current branch of the binary tree.

If the rest of the binary tree is explored, it eventually turns out that the best solution found so far (impute P and C) is also the optimal solution. A possible consistent record obtained by imputing P and C is: $(T, P, C, N) = (100, 40, 60, 5)$. This solution has the nice interpretation that the original values of *profit* and *total costs* were overstated by a factor of 1,000. It is of interest to note that, if only the hard edits are used in this example, then the first solution found above (impute only P) is the optimal solution. In that case, there is only one way to obtain a consistent record: $(T, P, C, N) = (100; -59{,}900; 60{,}000; 5)$. This illustrates that, in this example at least, soft edits are important for finding imputations that are not only consistent with the hard edits, but also plausible.

7 Application

To test the new error localisation algorithm in practice, a prototype implementation was written using the R programming language. This prototype draws heavily on the existing error localisation functionality in R that was made available in the `editrules` package (De Jonge and Van der Loo 2011; Van der Loo and De Jonge 2011).

To test the prototype, an artificial data set was constructed by selecting twelve numerical variables $(x_1, \ldots, x_{12})$ from the Dutch structural business statistics of 2007 for the wholesale sector. We selected all records pertaining to medium-sized businesses (with 10 to 100 employees) that had been edited manually during regular production, and divided these into two data sets of 728 records each. Both of the original data sets were considered error-free. We introduced a substantial number of random errors into one of the data sets by applying the following procedure:

- in 4% of the original non-zero values, two digits were interchanged;
- in 4% of the original non-zero values, a random digit was added;
- in 4% of the original non-zero values, a random digit was omitted;
- in 4% of the original non-zero values, a random digit was replaced by another digit;
- 4% of the original non-zero values were multiplied by 25;
- 4% of the original non-zero values were divided by 25 and rounded to the nearest integer;
- 6% of the original non-zero values were replaced by zero;
- 5% of the original zero values were replaced by random integers from {1; …; 1,000};
- 10% of the original values of x_{11} and x_{12} were multiplied by –1.

This procedure was carried out in such a way that at most one change could occur in each value. The second data set was left error-free and was used as reference data.

Table 7.1 shows the hard and soft edits that were applied to the test data. The hard edits were copied from the regular production system. The soft edits were identified by examining a number of univariate and bivariate distributions in the reference data.

Table 7.1
The edits that were used in the test application

hard edits:	$x_1 + x_2 = x_3$
	$x_2 = x_4$
	$x_5 + x_6 + x_7 = x_8$
	$x_3 + x_8 = x_9$
	$x_9 - x_{10} = x_{11}$
	$x_j \geq 0 \quad (j = 1,\ldots,10 \text{ and } j = 12)$
soft edits:	$x_2 \geq 0.5x_3$
	$x_3 \geq 0.9x_9$
	$x_5 + x_6 \geq x_7$
	$x_9 \geq 50x_{12}$
	$x_9 \leq 5,000x_{12}$
	$x_{11} \leq 0.4x_9$
	$x_{11} \geq -0.1x_9$
	$x_{12} \geq 1$
	$x_{12} \geq 5$
	$x_{12} \leq 100$

The error localisation algorithm was applied to the data set with artificial errors using several different setups. Throughout, all confidence weights w_j^N were chosen equal to 1, and the parameter λ in (3.1) was chosen equal to $1/2$. We considered the following approaches:

A. The first test used only the hard edits from Table 7.1.
B. The second test used all edits from Table 7.1, with all edits interpreted as hard edits.
C. The third test used all edits from Table 7.1, with a distinction between hard and soft edits. Each soft edit received the same fixed failure weight $s_k = 1$.
D. The fourth test was similar to the third test, but with fixed failure weights that differed between soft edits. For each soft edit, s_k was calculated as the fraction of records in the reference data set that satisfied the edit. Thus, a soft edit received a lower failure weight if it was failed more often in the reference data set, and vice versa. The rationale behind this is that all soft edit failures occurring in the reference data were caused by unusual but correct combinations of values. By associating low weights to soft edits that are often failed in the reference data, we ensure that these edits may also be failed more easily when editing the test data.

Since the distribution of errors in our test data set was known, we could directly evaluate the performance of each automatic error localisation approach. To this end, we used several quality indicators. Consider the following 2×2 contingency table:

		detected : error	**no error**
true:	error	*TP*	*FN*
	no error	*FP*	*TN*

The first quality indicator measures the proportion of true errors that were missed by the algorithm (proportion of false negatives):

$$\alpha = \frac{FN}{TP + FN}.$$

The second quality indicator measures the proportion of correct values that were mistaken for errors by the algorithm (proportion of false positives):

$$\beta = \frac{FP}{FP + TN}.$$

The third quality indicator measures the overall proportion of wrong decisions made by the algorithm:

$$\gamma = \frac{FN + FP}{TP + FN + FP + TN}.$$

These three indicators evaluate the performance of the algorithm with respect to identifying individual values as correct or erroneous. They have been used in previous evaluation studies; see, for instance, Pannekoek and De Waal (2005).

To evaluate the performance of the algorithm from a slightly different angle, we also calculated the percentage of records for which the algorithm found exactly the right solution – that is, the solution that identifies as erroneous all erroneous values and only these. This indicator is denoted by δ. A good editing approach should have low scores on α, β, and γ, but a high score on δ.

Table 7.2 shows the values of the quality indicators for editing approaches A, B, C, and D. It can be seen that approach B is outperformed by the other approaches on all measures, except for the proportion of missed errors. Thus, using the soft edits as if they were hard edits does not work well for this data set; in fact, better results are achieved by approach A, which does not use the soft edits at all. It can also be seen that approaches C and D, which use the new algorithm to take the soft edits into account, yield better results than approaches A and B, which use the old algorithm. Overall, approach D appears to achieve the best results in this experiment. Compared with approach A, approach D in fact correctly identifies more errors *and* more correct values.

It should be noted that, under the old definition of the error localisation problem, approaches A and B represent the two extreme options available for using soft edits: either not using them, or using them all as hard edits. As a compromise between these options, one could also decide to use only a subset of the soft edits as hard edits and discard the others. We did not test this

approach during the experiment. One might expect that it would lead to scores on the $\alpha, \beta, \gamma,$ and δ measures in between those of approaches A and B.

Table 7.2
Results of automatic error localisation for the artificial data

approach	quality indicators α	β	γ	δ
A	0.364	0.047	0.115	40%
B	0.232	0.131	0.153	37%
C	0.227	0.060	0.096	47%
D	0.253	0.037	0.083	52%

8 Conclusion

In this article, we proposed a new formulation of the error localisation problem which can take the distinction between hard and soft edits into account. In addition, we showed that a modified version of the branch-and-bound algorithm of De Waal and Quere (2003) can be used to solve this new error localisation problem. It was suggested that this new algorithm can be used to increase the quality of automatic editing. This suggestion was confirmed by the empirical results reported in Section 7, although it should be stressed that these results were obtained with data containing synthetic errors. An application is currently being investigated of the new error localisation algorithm to realistic data.

It remains an open problem how the costs of soft edit failures may best be modelled, *i.e.*, how the term D_{soft} in (3.1) should be defined. The different results with approaches C and D in Section 7 demonstrate that the quality of automatic error localisation may be improved by a suitable choice of failure weights. It will be interesting to see to what extent the quality of automatic editing may be improved further by experimenting with different combinations of failure weights s_k, confidence weights w_j, and the balancing parameter λ in (3.1).

Other forms of D_{soft} than (3.2) could also be considered, including forms that depend on the sizes of the soft edit failures. As mentioned in Section 3, it is intuitively appealing to take the amounts by which soft edits are failed into account in the error localisation problem, so that larger soft edit failures yield higher values of D_{soft}. One interesting choice for D_{soft} could be the Mahalanobis distance of soft edit failures, as suggested by Hedlin (2003) in a different context. It should be noted that the algorithm from Section 6 may be used to solve the error localisation problem for all choices of D_{soft} that can be expressed as (reasonably well-behaved) functions of $z_1, \ldots, z_{K_S}$. One simply uses the appropriate expression for D_{soft} as the target function in problem (5.1). On the other hand, if D_{soft} depends explicitly on the sizes of the soft edit failures, then we have to resort to a more complex approach. In particular, the information provided by Theorems 1 and 2 is no longer sufficient, because we now need to know not only which soft edits will be failed after imputation but also the amounts by which they will be failed. An approach for solving the error localisation problem in this more complicated situation can be found in Scholtus (2011).

In summary, it remains to be seen how the theoretical results outlined in this article should be applied to obtain the best results in practice. Nevertheless, given that subject-matter experts use the conceptual difference between hard and soft edits during manual editing, it seems evident that the new error localisation algorithm has the potential to increase the quality of automatic editing.

Acknowledgements

The views expressed in this article are those of the author and do not necessarily reflect the policies of Statistics Netherlands. The author would like to thank: Jeroen Pannekoek and Ton de Waal for their many helpful suggestions; Edwin de Jonge and Mark van der Loo for programming the branch-and-bound algorithm in R; Sevinç Göksen for collecting the evaluation results presented in Section 7; and finally the Associate Editor and two anonymous referees for their constructive comments.

Appendix A Proofs

A.1 Proof of Theorem 1

In order to prove Theorem 1, it is convenient to prove first an auxiliary lemma. Suppose that Ψ_q is obtained from Ψ_{q-1} by eliminating x_g. We define, for each edit ψ_q^k, the index set A_q^k of the edit(s) in Ψ_{q-1} from which it has been derived. That is to say, we define $A_q^k := \{l\}$ if ψ_q^k is obtained by copying the edit ψ_{q-1}^l, and we define $A_q^k := \{s,t\}$ if ψ_q^k is obtained by eliminating a variable from the pair of edits $(\psi_{q-1}^s, \psi_{q-1}^t)$.

Lemma 1. Consider the situation of Theorem 1 for $q \geq 1$, and suppose that x_g has been eliminated to obtain Ψ_q from Ψ_{q-1}. Let A be a representing set of the index sets A_q^k belonging to all $\psi_q^k \in \Psi_q^{(2)}$. Then there exists a value for x_g that, together with the original values of the variables that are involved in Ψ_q, satisfies all edits in Ψ_{q-1} except those in A.

Proof (of Lemma 1). By construction, A contains all indices of failed edits from Ψ_{q-1} which do not involve x_g. Hence, the only way for the lemma to be false would be if there existed two edits that involve x_g, say ψ_{q-1}^s and ψ_{q-1}^t, with $s \notin A$ and $t \notin A$, so that it is not possible to find a value for x_g that satisfies both edits simultaneously. In this case, an implied edit in Ψ_q is generated by eliminating x_g from ψ_{q-1}^s and ψ_{q-1}^t. Moreover, by the fundamental property given at the end of Section 2.3, this implied edit must be failed by the original values of the other variables, *i.e.*, the implied edit must be an element of $\Psi_q^{(2)}$. But this would contradict the assumption that A is a representing set of A_q^k for all $\psi_q^k \in \Psi_q^{(2)}$. Hence, it is impossible to find such a pair of edits, and the lemma follows.

The proof of Theorem 1 now proceeds by induction on the number of treated variables q. For $q = 0$, the statement is trivial. For $q = 1$, the theorem follows as a special case of Lemma 1; note that $B_1^k \equiv A_1^k$. We suppose therefore that the statement has been proved for all $q \in \{0,1,\ldots,Q-1\}$, and we consider the case $q = Q$, with $Q \geq 2$.

If Ψ_Q is obtained from Ψ_{Q-1} by fixing a variable to its original value, and B is a representing set of the sets B_Q^k for the failed edits from Ψ_Q, then by construction B is also a representing set of the sets B_{Q-1}^k for the failed edits from Ψ_{Q-1}. Thus, in this case, the statement for $q = Q$ follows immediately from the induction hypothesis.

Hence, we are left with the case that Ψ_Q is obtained from Ψ_{Q-1} by eliminating a variable, say x_g. We define, for each $\psi_Q^k \in \Psi_Q^{(2)}$, the index set A_Q^k of the edit(s) from Ψ_{Q-1} from which ψ_Q^k is derived, as above. Next, we use B to construct a set A, by applying the following procedure to each $\psi_Q^k \in \Psi_Q^{(2)}$:

- If ψ_Q^k is obtained by copying ψ_{Q-1}^l (so $A_Q^k = \{l\}$ and $B_Q^k = B_{Q-1}^l$), then we add l to A.
- If ψ_Q^k is obtained by eliminating x_g from ψ_{Q-1}^s and ψ_{Q-1}^t (so $A_Q^k = \{s, t\}$ and $B_Q^k = B_{Q-1}^s \cup B_{Q-1}^t$), then we add s to A if B contains an element of B_{Q-1}^s, and we add t to A otherwise.

It is easy to see that this procedure produces a representing set A of the index sets A_Q^k for all $\psi_Q^k \in \Psi_Q^{(2)}$.

According to Lemma 1, there exists a value for x_g which, together with the original values of the $p - q$ variables that have not been treated, satisfies the edits in Ψ_{Q-1} except those in A. That is to say, Ψ_{Q-1} can be partitioned similarly to Ψ_Q as $\Psi_{Q-1} = \Psi_{Q-1}^{(1)} \cup \Psi_{Q-1}^{(2)}$, where $\Psi_{Q-1}^{(2)}$ contains the edits with indices in A. Moreover, it is not difficult to see that the above procedure implies that B is a representing set of the index sets B_{Q-1}^k for all $\psi_{Q-1}^k \in \Psi_{Q-1}^{(2)}$. Hence, the induction hypothesis establishes that, given the original values of the variables that have not been eliminated *and* given the chosen value for x_g, there exist values for the other eliminated variables that satisfy all the original edits except those in B. This shows that the statement holds for $q = Q$ and completes the proof of Theorem 1.

A.2 Proof of Theorem 2

To prove Theorem 2, we start again with an auxiliary lemma. Analogous to the numerical case, when Ψ_q is obtained from Ψ_{q-1} by eliminating v_g, we define the index set A_q^k of edits in Ψ_{q-1} from which the edit $\psi_q^k \in \Psi_q$ is derived. To be precise, we define $A_q^k := \{l\}$ if ψ_q^k is obtained by copying the edit ψ_{q-1}^l, and we define $A_q^k := T$ if ψ_q^k is obtained by eliminating a variable from the set of edits $\psi_{q-1}^t (t \in T)$.

Lemma 2. Consider the situation of Theorem 2 for $q \geq 1$, and suppose that v_g has been eliminated to obtain Ψ_q from Ψ_{q-1}. Let A be a representing set of the index sets A_q^k belonging to all $\psi_q^k \in \Psi_q^{(2)}$. Then there exists a value for v_g that, together with the original values of the variables that are involved in Ψ_q, satisfies all edits in Ψ_{q-1} except those in A.

Proof (of Lemma 2). By construction, A contains all indices of failed edits from Ψ_{q-1} which do not involve v_g. Hence, the only way for the lemma to be false would be if there existed edits that involve v_g, say $\psi_{q-1}^{t_1}, \ldots, \psi_{q-1}^{t_r}$, with $A \cap \{t_1, \ldots, t_r\} = \varnothing$, so that it is not possible to find a value for v_g that satisfies these edits simultaneously, given the values of the other variables. Clearly, this could only happen if $F_g^{t_1} \cup \cdots \cup F_g^{t_r} = D_g$, since otherwise any value for v_g outside $F_g^{t_1} \cup \cdots \cup F_g^{t_r}$ would work. We may assume without loss of generality that $T' = \{t_1, \ldots, t_r\}$ is a minimal set having this property. Furthermore, it must hold in this case that for all variables involved in Ψ_q, the original value of v_j is contained in all sets $F_j^{t_1}, \ldots, F_j^{t_r}$. In other words, T' must satisfy properties (2.8) and (2.9). This means that T' would generate an implied edit in Ψ_q which, by the fundamental property given at the end of Section 2.3, must be failed by the original values of the remaining variables. However, this would contradict the assumption that A is a representing set of A_q^k for all $\psi_q^k \in \Psi_q^{(2)}$. This completes the proof of Lemma 2.

The proof of Theorem 2 is now completely analogous to that of Theorem 1, with Lemma 2 taking the role of Lemma 1.

References

Bankier, M., Lachance, M. and Poirier, P. (2000). *2001 Canadian Census Minimum Change Donor Imputation Methodology*. Working Paper, UN/ECE Work Session on Statistical Data Editing, Cardiff.

Bankier, M., and Crowe, S. (2009). *Enhancements to the 2011 Canadian Census E&I System*. Working Paper, UN/ECE Work Session on Statistical Data Editing, Neuchâtel.

Boskovitz, A., Goré, R. and Wong, P. (2005). *Data Editing and Logic*. Working Paper, UN/ECE Work Session on Statistical Data Editing, Ottawa.

Bruni, R. (2004). Discrete models for data imputation. *Discrete Applied Mathematics*, 144, 59-69.

Bruni, R. (2005). Error correction for massive datasets. *Optimization Methods and Software*, 20, 297-316.

Casado Valero, C., Del Castillo Cuervo-Arango, F., Mateo Ayerra, J. and De Santos Ballesteros, A. (1996). *Quantitative Data Editing*: *Quadratic Programming Method*. Presented at the COMPSTAT 1996 Conference, Barcelona.

De Jonge, E., and Van der Loo, M. (2011). *Manipulation of Linear Edits and Error Localization with the Editrules Package*. Discussion Paper 201120, Statistics Netherlands, The Hague.

De Waal, T. (2003a). Solving the error localization problem by means of vertex generation. *Survey Methodology*, 29, 1, 71-79.

De Waal, T. (2003b). *Processing of Erroneous and Unsafe Data*. PhD Thesis, Erasmus University, Rotterdam.

De Waal, T. (2005). *SLICE 1.5: a Software Framework for Automatic Edit and Imputation*. Working Paper, UN/ECE Work Session on Statistical Data Editing, Ottawa.

De Waal, T., Pannekoek, J. and Scholtus, S. (2011). *Handbook of Statistical Data Editing and Imputation*. New Jersey: John Wiley & Sons, Inc.

De Waal, T., and Quere, R. (2003). A fast and simple algorithm for automatic editing of mixed data. *Journal of Official Statistics*, 19, 383-402.

Fellegi, I.P., and Holt, D. (1976). A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association*, 71, 17-35.

Garfinkel, R.S., Kunnathur, A.S. and Liepins, G.E. (1986). Optimal imputation of erroneous data: Categorical data, general edits. *Operations Research*, 34, 744-751.

Garfinkel, R.S., Kunnathur, A.S. and Liepins, G.E. (1988). Error localization for erroneous data: Continuous data, linear constraints. *SIAM Journal on Scientific and Statistical Computing*, 9, 922-931.

Hedlin, D. (2003). Score functions to reduce business survey editing at the U.K. Office for National Statistics. *Journal of Official Statistics*, 19, 177-199.

Kovar, J., and Whitridge, P. (1990). Generalized edit and imputation system; Overview and applications. *Revista Brasileira de Estadistica*, 51, 85-100.

Little, R.J.A., and Smith, P.J. (1987). Editing and imputation of quantitative survey data. *Journal of the American Statistical Association*, 82, 58-68.

Magnus, J.R., Van Tongeren, J.W. and De Vos, A.F. (2000). National accounts estimation using indicator ratios. *Review of Income and Wealth*, 46, 329-350.

Mirsky, L. (1971). *Transversal Theory*. New York: Academic Press, Inc.

Nemhauser, G.L., and Wolsey, L.A. (1988). *Integer and Combinatorial Optimization*. New York: John Wiley & Sons, Inc.

Pannekoek, J., and De Waal, T. (2005). Automatic edit and imputation for business surveys: The Dutch contribution to the EUREDIT project. *Journal of Official Statistics*, 21, 257-286.

Ragsdale, C.T., and McKeown, P.G. (1996). On solving the continuous data editing problem. *Computers & Operations Research*, 23, 263-273.

Riera-Ledesma, J., and Salazar-González, J.J. (2003). *New Algorithms for the Editing and Imputation Problem*. Working Paper, UN/ECE Work Session on Statistical Data Editing, Madrid.

Sande, G. (1978). *An Algorithm for the Fields to Impute Problems of Numerical and Coded Data*. Technical Report, Statistics Canada.

Schaffer, J. (1987). Procedure for solving the data-editing problem with both continuous and discrete data types. *Naval Research Logistics*, 34, 879-890.

Scholtus, S. (2011). *Automatic Editing with Soft Edits*. Discussion Paper 201130, Statistics Netherlands, The Hague.

Schrijver, A. (1986). *Theory of Linear and Integer Programming*. New York: John Wiley & Sons, Inc.

Todaro, T.A. (1999). *Overview and Evaluation of the AGGIES Automated Edit and Imputation System*. Working Paper, UN/ECE Work Session on Statistical Data Editing, Rome.

Van der Loo, M., and De Jonge, E. (2011). *Manipulation of Categorical Data Edits and Error Localization with the Editrules Package*. Discussion Paper 201129, Statistics Netherlands, The Hague.

Williams, H.P. (1986). Fourier's method of linear programming and its dual. *The American Mathematical Monthly*, 93, 681-695.

Winkler, W.E. (1995). *Editing Discrete Data*. Working Paper, UN/ECE Work Session on Statistical Data Editing, Athens.


ELECTRONIC PUBLICATIONS AVAILABLE AT
PUBLICATIONS ÉLECTRONIQUES DISPONIBLE À
www.statcan.gc.ca

Survey Methodology, June 2013
Vol. 39, No. 1, pp. 91-120
Statistics Canada, Catalogue No. 12-001-X

Sparse and efficient replication variance estimation for complex surveys

Jae Kwang Kim and Changbao Wu[1]

Abstract

It is routine practice for survey organizations to provide replication weights as part of survey data files. These replication weights are meant to produce valid and efficient variance estimates for a variety of estimators in a simple and systematic manner. Most existing methods for constructing replication weights, however, are only valid for specific sampling designs and typically require a very large number of replicates. In this paper we first show how to produce replication weights based on the method outlined in Fay (1984) such that the resulting replication variance estimator is algebraically equivalent to the fully efficient linearization variance estimator for any given sampling design. We then propose a novel weight-calibration method to simultaneously achieve efficiency and sparsity in the sense that a small number of sets of replication weights can produce valid and efficient replication variance estimators for key population parameters. Our proposed method can be used in conjunction with existing resampling techniques for large-scale complex surveys. Validity of the proposed methods and extensions to some balanced sampling designs are also discussed. Simulation results showed that our proposed variance estimators perform very well in tracking coverage probabilities of confidence intervals. Our proposed strategies will likely have impact on how public-use survey data files are produced and how these data sets are analyzed.

Key Words: Bootstrap; Calibration; Jackknife; Linearization method; Replication weights; Sampling design; Spectral decomposition.

1 Introduction

Variance estimation is an important practical problem in sample surveys. In addition to analytic use of variances such as testing statistical hypotheses and constructing confidence intervals, variance estimation can also be used to provide descriptive measures on the accuracy of survey estimates and the efficiency of the given sampling design. There are two types of commonly used techniques for variance estimation under the design-based framework. The first is called the linearization method, which uses the standard variance formula applied either directly to the estimator if the parameter is a population total or to the linearized one-step Taylor

1. Jae Kwang Kim, Department of Statistics, Iowa State University, Ames IA 50011-1210. E-mail: jkim@iastate.edu; Changbao Wu, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo ON N2L 3G1. E-mail: cbwu@uwaterloo.ca.

series expansion of the estimator if the parameter is a nonlinear function of one or several population totals. The second is called the replication method, which constructs variance estimators in a simple systematic way using multiple sets of replication weights along with the original survey data set.

Replication variance estimation techniques have become very popular for design-based inferences using complex survey data. Some early practices using replication weights go back to 1970s at the U.S. Bureau of the Census, Bureau of Labor Statistics and Westat (Dippo, Fay and Morganstein 1984). It is now a routine practice for survey organizations to provide replication weights together with survey data. The most attractive feature of this approach is that it works the same way regardless of the complexity of the parameter. For parameters that are smooth functions of population means or totals, the "linearization" step has been automatically built into the estimation process and computation of partial derivatives involved in the Taylor series expansion is not required. It is extremely user-friendly for multi-purpose data analyses once the survey data set is released together with replication weights. Furthermore, the use of replication methods reduces concerns on confidentiality issues since detailed design information such as stratum or cluster identifier is not released (Lu and Sitter 2008).

Replication weights are typically constructed by the bootstrap, the jackknife or the balanced repeated replication (BRR) methods. Rust and Rao (1996), Shao (1996, 2003) and Wolter (2007) provided excellent overviews on the topic. There are three major issues in the construction of replication weights: *validity*, *efficiency* and *sparsity*. Validity refers to the asymptotic unbiasedness of replication variance estimators under the given sampling design. The asymptotic unbiasedness of an estimator is generally a weaker concept than the estimator being consistent. If the coefficient of variation of the variance estimator goes to zero, then the asymptotically unbiased variance estimator is also consistent. Efficiency is measured by the relative performance of the replication variance estimator to the standard linearization variance estimator which is

viewed as fully efficient. Sparsity refers to the number of sets of replication weights required to achieve fully efficient variance estimation.

Validity of replication variance estimators was discussed by Krewski and Rao (1981), Shao and Tu (1995) and Fuller (2009a), among others. Efficiency and stability of replication variance estimators were discussed by Rust and Kalton (1987) and Jang and Eltinge (2009). For sparsity, Kott (2001) considered using delete-a-group jackknife to achieve sparsity under certain designs, and Lu, Brick and Sitter (2006) also discussed combining strata for sparse replication variance estimation.

Most replication methods discussed in the literature are only valid for certain sampling designs. For example, the jackknife method is commonly used for stratified random sampling (Krewski and Rao 1981). The bootstrap method has several popular procedures, including the without-replacement bootstrap method (Gross 1980; McCarthy and Snowden 1985), the re-scaling bootstrap method (Rao and Wu 1988; Preston 2009) and the mirror-match bootstrap method (Sitter 1992). These procedures, however, are only applicable for certain types of sampling designs.

The sparsity of a replication method depends on how the replication weights are constructed. The number of sets of the jackknife replication weights is related to the number of units in the sample and can be very large if the sample size is large. Bootstrap methods typically require at least 1,000 sets of replication weights in order to achieve the desired level of efficiency. As a compromise, most survey organizations provide 500 sets of bootstrap weights alongside the main survey variables. The resulting data sets are still too big for data users to have visual checks and can be very cumbersome to manipulate in practice.

This paper presents methods for constructing efficient and sparse replication weights for variance estimation under the design-based framework. By maintaining full efficiency of the resulting variance estimator for key variables with a smaller number of sets of replication weights, our methods address one of the major tasks at the data file preparation stage and can

easily be applied by survey runners to reduce the burden of data users in dealing with excessively large data files. A major limitation of our proposed method is that it does not directly handle situations where design weights are adjusted for nonresponse or calibrated to known auxiliary population information.

In Section 2, we present a general procedure for constructing replication weights based on the method of Fay (1984) and Fay and Dippo (1989), which provides fully efficient replication weights for arbitrary sampling designs. In Section 3, we discuss two strategies, random sampling and calibration weighting, for constructing sparse replication weights. By using a novel application of the calibration technique, our proposed methods allow the use of a small number of sets of replication weights while the resulting replication variance estimators remain efficient. In Section 4, some asymptotic theory for the validity of the replication variance estimator is presented. In Section 5, extensions to some balanced sampling designs are discussed. In Section 6, we report results from a simulation study, using real data from Statistics Canada's Family Expenditure Survey, to evaluate the effectiveness of the proposed strategies for replication variance estimation. Some concluding remarks are given in Section 7.

2 A general procedure for constructing fully efficient replication weights

In principle, we can construct replication weights for any measurable sampling design, using the method outlined in Fay (1984) and Fay and Dippo (1989), such that the resulting replication variance estimators are algebraically equivalent to the standard linearization variance estimators.

Let $\mathcal{U} = \{1, 2, \ldots, N\}$ be the set of N units in the finite population and $\mathcal{S} = \{1, 2, \ldots, n\}$ be the set of n units in the sample, selected according to a probability sampling design. Let $w_i = 1 / \pi_i$ be the basic design weight, where $\pi_i = P(i \in \mathcal{S})$ is the first order inclusion probability for unit i.

Let y_i be the value of the study variable y for unit i and $t_y = \sum_{i=1}^{N} y_i$ be the population total of interest. The Horvitz-Thompson estimator of t_y is given by

$$\hat{t}_y = \sum_{i \in \mathcal{S}} w_i y_i. \tag{2.1}$$

The estimator $\hat{t}_y$ given in (2.1) is also called the expansion estimator, with the basic design weight w_i denoting the number of units in the population represented by unit i in the sample. The standard variance estimator of $\hat{t}_y$ can be written as

$$v = \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}} \Omega_{ij} y_i y_j, \tag{2.2}$$

where $\Omega_{ij} = (\pi_{ij} - \pi_i \pi_j) / (\pi_{ij} \pi_i \pi_j)$ and $\pi_{ij} = P(i, j \in \mathcal{S})$ is the second order joint inclusion probability for (ij). It is assumed that $\pi_{ij} > 0$ for all (ij). Note that $\pi_{ii} = \pi_i$. The standard variance estimator v is often viewed as fully efficient since it is the Horvitz-Thompson estimator of the design-based variance $V(\hat{t}_y)$.

Let $\boldsymbol{\Delta} = (\Omega_{ij})$ be an $n \times n$ matrix. We can re-write (2.2) as $v = \boldsymbol{y}' \boldsymbol{\Delta} \boldsymbol{y}$, where $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$ is the vector of sampled y_i's. The matrix $\boldsymbol{\Delta}$ is nonnegative definite and can be decomposed as

$$\boldsymbol{\Delta} = \sum_{k=1}^{p} \lambda_k \boldsymbol{\delta}_k \boldsymbol{\delta}_k' \tag{2.3}$$

for some $\lambda_k > 0$ and some n-dimensional vectors $\boldsymbol{\delta}_k, k = 1, 2, \ldots, p$. The most well-known decomposition (2.3) is given by the spectral decomposition where $\boldsymbol{\delta}_k$ is the eigenvector associated with the eigenvalue λ_k. In practice, very small eigenvalues are often ignored for computational reasons. For stratified sampling, the matrix $\boldsymbol{\Delta}$ is block-diagonal so the computational burden may be alleviated. However, we do not restrict (2.3) to the spectral decomposition. Any decomposition satisfying (2.3) can be used.

Suppose that we want to express the fully efficient variance estimator v given by (2.2) as a replication variance estimator in the form of

$$v_R = \sum_{k=1}^{L} c_k (\hat{t}_y^{(k)} - \hat{t}_y)^2, \tag{2.4}$$

where $\hat{t}_y^{(k)} = \sum_{i \in \mathcal{S}} w_i^{(k)} y_i$, $\boldsymbol{w}^{(k)} = (w_1^{(k)}, \ldots, w_n^{(k)})'$ is the k^{th} set of replication weights, $c_k > 0$ is the factor associated with the k^{th} set of replication weights and L is the total number of replications; see Kim, Navarro and Fuller (2006) for further discussion.

The form given by (2.4) does not include all replication variance estimators. For instance, Campbell (1980) provided a jackknife variance estimator where the pseudovalues are derived based on the von Mises approximation to the parameter of interest. Nevertheless, most replication variance estimators can be put in this form.

We have the following result on the construction of $\boldsymbol{w}^{(k)}$ for v_R based on the decomposition (2.3).

Theorem 1. The fully efficient variance estimator v and the replication variance estimator v_R are algebraically identical if we let $L = p$ and $\boldsymbol{w}^{(k)} = \boldsymbol{w} + (\lambda_k / c_k)^{1/2} \boldsymbol{\delta}_k$, where $\boldsymbol{w} = (w_1, \ldots, w_n)'$ is the set of original basic design weights.

Proof. The proof follows directly from the fact that $v = \boldsymbol{y}'\boldsymbol{\Delta}\boldsymbol{y} = \sum_{k=1}^{p} \lambda_k (\boldsymbol{\delta}_k' \boldsymbol{y})^2$ and that $\hat{t}_y^{(k)} - \hat{t}_y = (\boldsymbol{w}^{(k)} - \boldsymbol{w})'\boldsymbol{y} = (\lambda_k / c_k)^{1/2} \boldsymbol{\delta}_k' \boldsymbol{y}$.

The choices of c_k's can be arbitrary and bear no impact on the validity and efficiency of the replication variance estimators. However, certain choices of c_k will result in replication weights with negative values, which is undesirable as it may produce negative replicates for the parameters that are always positive. In practical situations one can always choose relatively large c_k to avoid negative values for replication weights. In our simulation study (*Case I*) reported in Section 5, the problem of negative replication weights can be eliminated with the choice of $c_k = 1$.

The replication variance estimator $v_R = \sum_{k=1}^{L} c_k(\hat{t}_y^{(k)} - \hat{t}_y)^2$ with $L = p$ and replication weights $\boldsymbol{w}^{(k)} = \boldsymbol{w} + (\lambda_k / c_k)^{1/2}\boldsymbol{\delta}_k$ is fully efficient for an arbitrary variable y. It also provides fully efficient variance estimator for $\hat{\theta}$ when θ is a smooth function of population means or totals. Practical implementation of the method depends crucially on two related issues: (i) the feasibility of the decomposition of the $n \times n$ matrix $\boldsymbol{\Delta}$ specified in (2.3); and (ii) the number of sets of replication weights required to achieve the full efficiency determined by $p = \text{rank}(\boldsymbol{\Delta})$.

As for the first issue, modern advances in computational power and improved performances of available software packages make it possible to do the spectral decomposition with relatively large n. For instance, on a 12-CPU unix machine with 96 gigabytes of memory, the R function *eigen*() can handle matrices of sizes at least as large as $n = 4{,}000$. Note that the computational task involved here is for survey runners at the data preparation stage and is not for users of the data files. As for the second issue, the value of p is related to the given sampling design. For simple random sampling without replacement, we have

$$\boldsymbol{\Delta} = N^2(1 - n / N)(n(n-1))^{-1}(\boldsymbol{I}_n - \mathbf{1}_n\mathbf{1}_n' / n),$$

where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix and $\mathbf{1}_n = (1, 1, \ldots, 1)'$ is the $n \times 1$ vector of 1's. It follows that $p = \text{rank}(\Delta) = \text{trace}(\boldsymbol{I}_n - \mathbf{1}_n\mathbf{1}_n' / n) = n - 1$. This is typically the case for single stage unequal probability sampling designs. For stratified simple random sampling, we have $p = n - H$, where H is the total number of strata.

It should be noted that $p \leq n$ for any sampling design and the exact value of p is not required for the proposed procedure to be implemented. However, since the values of p and n have the same order of magnitude, the proposed method requires a large number of replicates whenever n is large. Under the current practices in sample surveys, the fully efficient replication weights described above become immediately implementable if $p \leq 500$ and the second order inclusion probabilities π_{ij} are available. When p is large, a two-stage procedure to be described

in Section 3 can be used to produce a small number L_0 sets of replication weights for public-use data files.

In some cases, the spectral decomposition (2.3) can be avoided. For example, Deville (1999) argued that the variance estimator of $\hat{t}_y$ under unequal probability sampling designs with fixed sample size can be approximated by

$$v \doteq c\sum_{i\in\mathcal{S}}(1-\pi_i)\left(\frac{y_i}{\pi_i}-\tilde{t}_y\right)^2 \tag{2.5}$$

where $c=\left(1-\sum_{i\in\mathcal{S}}b_i^2\right)^{-1}, b_i=(1-\pi_i)\big/\sum_{k\in\mathcal{S}}(1-\pi_k)$ and $\tilde{t}_y=\sum_{i\in\mathcal{S}}b_i(y_i/\pi_i)$. More generally, we consider the following form of matrix $\boldsymbol{\Delta}$ in $v=\boldsymbol{y}'\boldsymbol{\Delta}\boldsymbol{y}$, where

$$\boldsymbol{\Delta}=\boldsymbol{\Delta}_0-\boldsymbol{\Delta}_0\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Delta}_0\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Delta}_0 \tag{2.6}$$

where $\boldsymbol{\Delta}_0=\text{diag}\{\lambda_1,\ldots,\lambda_n\}, \lambda_i>0$ for all $i=1,2,\ldots,n, \boldsymbol{X}'=(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)$ and $\boldsymbol{x}_i$ is a vector of design and auxiliary variables. Many elementary single-stage sampling designs take the form (2.6) for variance estimation. In particular, Deville's formula in (2.5) can be expressed as $v\doteq\boldsymbol{y}'\boldsymbol{\Delta}\boldsymbol{y}$ with $\boldsymbol{\Delta}$ given by (2.6), where $\lambda_i=c\pi_i^{-2}(1-\pi_i)$ in $\boldsymbol{\Delta}_0$ and $\boldsymbol{x}_i=\pi_i$. The conditional Poisson sampling design to be discussed in Section 5 also takes the form (2.6) where $\boldsymbol{x}_i$ are the design variables in the design constraint $\sum_{i\in\mathcal{S}}\pi_i^{-1}\boldsymbol{x}_i=\sum_{i=1}^N\boldsymbol{x}_i$.

For the matrix given by (2.6), it can be shown that

$$\boldsymbol{y}'\boldsymbol{\Delta}\boldsymbol{y}=(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})'\boldsymbol{\Delta}_0(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}),$$

where $\hat{\boldsymbol{\beta}}=(\boldsymbol{X}'\boldsymbol{\Delta}_0\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Delta}_0\boldsymbol{y}$. Thus, we have

$$\boldsymbol{y}'\boldsymbol{\Delta}\boldsymbol{y}=\sum_{k=1}^n\lambda_k(y_k-\boldsymbol{x}_k'\hat{\boldsymbol{\beta}})^2, \tag{2.7}$$

which is useful in deriving an expression for replication variance estimator in the form given by (2.4). The fully efficient variance estimator v in (2.7) and the replication variance estimator v_R

in (2.4) are algebraically identical if we let $L = n$ and $\boldsymbol{w}^{(k)} = \boldsymbol{w} + (\lambda_k / c_k)^{1/2}\boldsymbol{\delta}_k$, where $\boldsymbol{w} = (w_1, \ldots, w_n)'$ is the set of original basic design weights and $\boldsymbol{\delta}_k = (\delta_{1k}, \ldots, \delta_{nk})'$ with

$$\delta_{ik} = \begin{cases} -1 + \boldsymbol{x}_k'(\boldsymbol{X}'\boldsymbol{\Delta}_0\boldsymbol{X})^{-1}\boldsymbol{x}_i\lambda_i & \text{if } i = k \\ \boldsymbol{x}_k'(\boldsymbol{X}'\boldsymbol{\Delta}_0\boldsymbol{X})^{-1}\boldsymbol{x}_i\lambda_i & \text{otherwise.} \end{cases}$$

The proof follows directly from the fact that $\boldsymbol{\delta}_k'\boldsymbol{y} = -y_k + \boldsymbol{x}_k'\hat{\boldsymbol{\beta}}$, $\boldsymbol{y}'\boldsymbol{\Delta}\boldsymbol{y} = \sum_{k=1}^{n}\lambda_k(\boldsymbol{\delta}_k'\boldsymbol{y})^2$ and that $\hat{t}_y^{(k)} - \hat{t}_y = (\boldsymbol{w}^{(k)} - \boldsymbol{w})'\boldsymbol{y} = (\lambda_k / c_k)^{1/2}\boldsymbol{\delta}_k'\boldsymbol{y}$.

3 Sparse and efficient replication weights

Large-scale complex surveys usually have a relatively large sample size ranging from a few hundreds to many thousands. The fully efficient replication weights described in Section 2 or replication weights constructed by some existing methods such as the jackknife or the bootstrap methods would involve a very large number of sets of weights. Although valid replication weights provide enormous convenience to the users of survey data, who are not necessarily the survey runners, the burden of manipulating a data set with hundreds or even thousands of replicate weights can be enormous. As a result, how to achieve efficient replication variance estimation with a relatively small number of replicate weights is a question with both theoretical and practical value.

We propose two strategies to construct sparse and efficient replication weights. We start with a large number L sets of replication weights. These initial weights may be produced using the general method described in Section 2 or by existing methods. Suppose they can be viewed as fully efficient. The first strategy is to select a small number L_0 sets of weights from the initial large number L sets of weights using a probability sampling method. The small number L_0 satisfies the desired *sparsity* and the random selection procedure guarantees *validity* of the resulting variance estimators. The second strategy is to achieve *efficiency* through a novel

weight-calibration procedure. The L_0 sets of calibrated replication weights provide fully efficient variance estimators for variables used in the calibration and also highly efficient variance estimators for variables related to calibration variables.

3.1 Achieve sparsity and efficiency through random sampling

Suppose that the fully efficient replication variance estimator is given by $v_R = \sum_{k=1}^{L} c_k (\hat{t}_y^{(k)} - \hat{t}_y)^2$, with replication weights constructed by using Theorem 1. Observe that v_R can be viewed as a finite population total. If we want to use $L_0 (< L)$ sets of replication weights to provide valid inference for variance estimation, the following simple strategy can be used. First, select L_0 sets of weights from the original L sets of weights by simple random sampling without replacement. For notational simplicity and without loss of generality, we denote the selected sets of weights by $\boldsymbol{w}^{(j)}$, $j = 1, 2, \ldots, L_0$. Then, calculate the replication variance estimator of $\hat{t}_y$ based on the L_0 sets of weights as

$$v_R^{(1)} = \frac{L}{L_0} \sum_{j=1}^{L_0} c_j (\hat{t}_y^{(j)} - \hat{t}_y)^2. \tag{3.1}$$

The variance estimator $v_R^{(1)}$ is still unbiased for an arbitrary variable y, since $E^*(v_R^{(1)}) = \sum_{k=1}^{L} c_k (\hat{t}_y^{(k)} - \hat{t}_y)^2 = v_R$, where $E^*(\cdot)$ denotes the expectation under the random selection of L_0 sets of weights.

An alternative form of the replication variance estimator based on the L_0 sets of weights can be derived as follows. Noting that $\hat{t}_y^{(k)} - \hat{t}_y = (\lambda_k / c_k)^{1/2} \boldsymbol{\delta}_k' \boldsymbol{y}$, we can re-write the fully efficient variance estimator as

$$v_R = m \left\{ \sum_{k=1}^{L} \lambda_k (\boldsymbol{\delta}_k' \boldsymbol{y})^2 \Big/ \sum_{k=1}^{L} \lambda_k \right\},$$

where $m = \sum_{k=1}^{L} \lambda_k$. The $\boldsymbol{\delta}_k$'s are orthogonal eigenvectors satisfying $|\boldsymbol{\delta}_k| = 1$ under spectral decomposition and $\boldsymbol{\delta}_k' \boldsymbol{y}$ are projections of $\boldsymbol{y}$ onto the n-dimensional unit-sphere. It is very

natural to use the following weighted version for the variance estimator of $\hat{t}_y$ based on the L_0 randomly selected sets of weights:

$$v_R^{(2)} = m\left\{\sum_{j=1}^{L_0}\lambda_j(\boldsymbol{\delta}_j'\boldsymbol{y})^2 \Big/ \sum_{j=1}^{L_0}\lambda_j\right\} = \frac{m}{m_0}\sum_{j=1}^{L_0}c_j(\hat{t}_y^{(j)} - \hat{t}_y)^2, \tag{3.2}$$

where $m_0 = \sum_{j=1}^{L_0}\lambda_j$. Noting that $v_R^{(2)}$ is a ratio estimator of v_R, it is usually more efficient than $v_R^{(1)}$.

A third version of the replication variance estimator can be constructed by first selecting L_0 sets of weights with unequal probabilities and then using a Horvitz-Thompson estimator of v_R. Note that we can view $v_R = \sum_{k=1}^{L}\lambda_k(\boldsymbol{\delta}_k'\boldsymbol{y})^2$ as a population total and λ_k as a size variable. We select L_0 sets of weights from the original L sets of weights with inclusion probabilities η_k proportional to λ_k. The resulting variance estimator of $\hat{t}_y$ is given by

$$v_R^{(3)} = \sum_{j=1}^{L_0}\frac{c_j}{\eta_j}(\hat{t}_y^{(j)} - \hat{t}_y)^2, \tag{3.3}$$

where $\eta_j = L_0\lambda_j \Big/ \sum_{k=1}^{L}\lambda_k$. It turns out that the eigenvalues λ_k differ substantially in magnitude and using λ_k as size measure leads to a large portion of the L sets of weights being selected with probability one. In the simulation study described in Section 5, we also included a fourth version of the replication variance estimator, denoted as $v_R^{(4)}$, with $\lambda_k^{1/2}$ as the size measure and $\eta_j = L_0\lambda_j^{1/2} \Big/ \sum_{k=1}^{L}\lambda_k^{1/2}$.

Another possible version of the replication variance estimator is to simply select L_0 sets of weights corresponding to the L_0 largest values of λ_k and then use $v_R^{(2)}$. Simulation results, not reported here, showed that the resulting variance estimator is severely biased and shouldn't be used in practice.

3.2 Achieve sparsity and efficiency through weight-calibration

We now discuss a novel approach of achieving sparsity without losing the efficiency of the variance estimators for some key variables. Suppose L_0 is the desired replication size, which is

much smaller than the sample size n. For example, the Natural Resources Inventory Survey (sponsored by the US department of Agriculture) used $L_0 = 29$ while the PSU sample size can be as large as $n = 300{,}000$. We present a weight-calibration technique that not only allows the use of a small L_0, but also provides fully efficient variance estimators for key population parameters. Our proposed strategy for constructing the smaller number L_0 sets of replication weights consists of the following four steps:

Step 1. Choose a set of key variables for which full efficiency of the variance estimator is desired. Let $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{im})'$ be the vector of key variables for the i^{th} unit included in the survey data file, where $m \leq L_0$. Among them can be important auxiliary variables and study variables as well as design variables. Let $\hat{t}_{\boldsymbol{z}} = \sum_{i \in \mathcal{S}} w_i \boldsymbol{z}_i$. Let $v_1(\hat{t}_{\boldsymbol{z}})$ be an $m \times m$ estimated variance-covariance matrix for $\hat{t}_{\boldsymbol{z}}$ computed by the standard linearization method or by a replication method that is fully efficient.

Step 2. Construct an initial L_0 sets of replication weights that produce an asymptotically unbiased variance estimator. These initial replicates can be obtained by a bootstrap method with L_0 replicates, or by the delete-a-group jackknife method of Kott (2001), or by the sampling method described in Section 3.1. Let $\boldsymbol{w}_0^{(k)} = (w_{10}^{(k)}, \ldots, w_{n0}^{(k)})', k = 1, \ldots, L_0$ be the initial sets of weights. Denote $\hat{t}_{y0}^{(k)} = \sum_{i \in \mathcal{S}} w_{i0}^{(k)} y_i$ and let

$$v_0 = \sum_{k=1}^{L_0} c_{k0} (\hat{t}_{y0}^{(k)} - \hat{t}_y)^2 \tag{3.4}$$

be the replication variance estimator based on the L_0 sets of weights.

We can apply the variance formula (3.4) to the vector of key variables $\boldsymbol{z}$ to get $v_0(\hat{t}_{\boldsymbol{z}}) = \sum_{k=1}^{L_0} c_{k0} (\hat{t}_{\boldsymbol{z}0}^{(k)} - \hat{t}_{\boldsymbol{z}})(\hat{t}_{\boldsymbol{z}0}^{(k)} - \hat{t}_{\boldsymbol{z}})'$, where $\hat{t}_{\boldsymbol{z}0}^{(k)} = \sum_{i \in \mathcal{S}} w_{i0}^{(k)} \boldsymbol{z}_i$. Note that $v_0(\hat{t}_{\boldsymbol{z}})$ is not as efficient as $v_1(\hat{t}_{\boldsymbol{z}})$ obtained in *Step* 1.

Step 3. Decompose the nonnegative definite variance-covariance matrix $v_1(\hat{t}_{\boldsymbol{z}})$ as

$$v_1(\hat{t}_{\boldsymbol{z}}) = \sum_{k=1}^{m} \alpha_k \boldsymbol{q}_k \boldsymbol{q}_k' \tag{3.5}$$

using the spectral decomposition or any other suitable methods. Let $\alpha_k = 0$ for $k = m+1, \ldots, L_0$ and define

$$\hat{t}_{\boldsymbol{z}}^{(k)} = \hat{t}_{\boldsymbol{z}} + (\alpha_k / c_{k0})^{1/2} \boldsymbol{q}_k, k = 1, 2, \ldots, L_0.$$

It follows that the L_0 pseudo-replicates $\hat{t}_{\boldsymbol{z}}^{(k)}$ defined above satisfy

$$\sum_{k=1}^{L_0} c_{k0} (\hat{t}_{\boldsymbol{z}}^{(k)} - \hat{t}_{\boldsymbol{z}})(\hat{t}_{\boldsymbol{z}}^{(k)} - \hat{t}_{\boldsymbol{z}})' = v_1(\hat{t}_{\boldsymbol{z}}), \tag{3.6}$$

due to the decomposition to $v_1(\hat{t}_{\boldsymbol{z}})$ given in (3.5). It should be noted that (3.5) bears no relation to the decomposition to $\boldsymbol{\Delta}$ described in Section 2 and the condition $m \leq L_0$ is required to make (3.6) possible.

Step 4. Improve the efficiency of v_0 computed from (3.4) for an arbitrary y variable through a weight-calibration procedure. For the k^{th} set of initial weights $\boldsymbol{w}_0^{(k)} = (w_{10}^{(k)}, \ldots, w_{n0}^{(k)})'$, the calibrated weights $\boldsymbol{w}_c^{(k)} = (w_{1c}^{(k)}, \ldots, w_{nc}^{(k)})'$ minimize the chi-square distance measure

$$\Phi(\boldsymbol{w}_c^{(k)}, \boldsymbol{w}_0^{(k)}) = \sum_{i \in \mathcal{S}_k} \tau_i (w_{ic}^{(k)} - w_{i0}^{(k)})^2 / w_{i0}^{(k)} \tag{3.7}$$

subject to the constraint

$$\sum_{i \in \mathcal{S}} w_{ic}^{(k)} \boldsymbol{z}_i = \hat{t}_{\boldsymbol{z}}^{(k)}, \tag{3.8}$$

where $\mathcal{S}_k = \left\{ i \middle| i \in \mathcal{S}; w_{i0}^{(k)} > 0 \right\}$, the τ_i's are known constants, and $\hat{t}_{\boldsymbol{z}}^{(k)}$ is the k^{th} pseudo replicate of $\hat{t}_{\boldsymbol{z}}$ defined in *Step* 3.

The calibrated weights $\boldsymbol{w}_c^{(k)} = (w_{1c}^{(k)}, \ldots, w_{nc}^{(k)})', k = 1, 2, \ldots, L_0$ are used in (3.4) to compute the final replication variance estimator $v_C(\hat{t}_y) = \sum_{k=1}^{L_0} c_{k0} (\hat{t}_{yc}^{(k)} - \hat{t}_y)^2$, where $\hat{t}_{yc}^{(k)} = \sum_{i \in \mathcal{S}} w_{ic}^{(k)} y_i$.

The proposed weight-calibration procedure ensures that the replication estimator $v_C(\hat{t}_{\mathbf{z}})$ based on the L_0 sets of calibrated weights matches exactly the fully efficient estimator $v_1(\hat{t}_{\mathbf{z}})$, due to the calibration constraints (3.8) and the equation (3.6). Furthermore, the calibrated replication weights provide more efficient variance estimators for an arbitrary y that is related to $\mathbf{z}$. To see this, we re-write $\hat{t}_{y0}^{(k)} = \sum_{i\in\mathcal{S}} w_{i0}^{(k)} y_i$ as

$$\hat{t}_{y0}^{(k)} = \hat{t}_{e0}^{(k)} + (\hat{t}_{\mathbf{z}0}^{(k)})'\hat{\boldsymbol{\beta}}, \tag{3.9}$$

where $\hat{t}_{e0}^{(k)} = \sum_{i\in\mathcal{S}} w_{i0}^{(k)} \hat{e}_i, \hat{e}_i = y_i - \mathbf{z}_i'\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}} = \left\{\sum_{i\in\mathcal{S}} w_i \mathbf{z}_i \mathbf{z}_i' / \tau_i\right\}^{-1} \sum_{i\in\mathcal{S}} w_i \mathbf{z}_i y_i / \tau_i$. Let $\hat{t}_e = \sum_{i\in\mathcal{S}} w_i \hat{e}_i$. The variance estimator of $\hat{t}_y$ based on the initial L_0 sets of weights can be expressed as

$$\begin{aligned} v_0(\hat{t}_y) &= \sum_{k=1}^{L_0} c_{k0} (\hat{t}_{y0}^{(k)} - \hat{t}_y)^2 \\ &= \sum_{k=1}^{L_0} c_{k0} (\hat{t}_{e0}^{(k)} - \hat{t}_e)^2 + \sum_{k=1}^{L_0} c_{k0} ((\hat{t}_{\mathbf{z}0}^{(k)})'\hat{\boldsymbol{\beta}} - (\hat{t}_{\mathbf{z}})'\hat{\boldsymbol{\beta}})^2 + 2\sum_{k=1}^{L_0} c_{k0} (\hat{t}_{e0}^{(k)} - \hat{t}_e)((\hat{t}_{\mathbf{z}0}^{(k)})'\hat{\boldsymbol{\beta}} - (\hat{t}_{\mathbf{z}})'\hat{\boldsymbol{\beta}}) \\ &= v_0(\hat{t}_e) + \hat{\boldsymbol{\beta}}' v_0(\hat{t}_{\mathbf{z}})\hat{\boldsymbol{\beta}} + 2\hat{\boldsymbol{\beta}}' \operatorname{cov}_0(\hat{t}_e, \hat{t}_{\mathbf{z}}), \end{aligned}$$

where $\operatorname{cov}_0(\hat{t}_e, \hat{t}_{\mathbf{z}})$ is the estimated covariance between $\hat{t}_e$ and $\hat{t}_{\mathbf{z}}$ based on the initial L_0 sets of replication weights. In many designs, we can choose a suitable τ_i such that $\operatorname{Cov}(\hat{t}_e, \hat{t}_{\mathbf{z}}) \doteq \mathbf{0}$. This is the case, for instance, with the choice of $\tau_i = (w_i - 1)^{-1}$ or $\tau_i = w_i^{-1}$ under Poisson sampling. Fuller (1998) discussed the required conditions in the context of two-phase sampling. It follows that

$$v_0(\hat{t}_y) \doteq v_0(\hat{t}_e) + \hat{\boldsymbol{\beta}}' v_0(\hat{t}_{\mathbf{z}})\hat{\boldsymbol{\beta}}. \tag{3.10}$$

Using similar argument, it can be shown that the variance estimator based on the L_0 sets of calibrated weights satisfies

$$v_C(\hat{t}_y) \doteq v_0(\hat{t}_e) + \hat{\boldsymbol{\beta}}' v_1(\hat{t}_{\mathbf{z}})\hat{\boldsymbol{\beta}}. \tag{3.11}$$

The variance estimator $v_C(\hat{t}_y)$ given by (3.11) is generally more efficient than $v_0(\hat{t}_y)$ given by (3.10), due to the use of $v_1(\hat{t}_{\mathbf{z}})$ instead of $v_0(\hat{t}_{\mathbf{z}})$. The gain of efficiency depends on the

relative magnitude of $\hat{\boldsymbol{\beta}}' v_1(\hat{t}_{\boldsymbol{z}})\hat{\boldsymbol{\beta}}$ over $v_0(\hat{t}_e)$. If y is highly correlated with $\hat{y} = \boldsymbol{z}'\hat{\boldsymbol{\beta}}$, the variance of the residual term $v_0(\hat{t}_e)$ will be relatively small. In this case $v_C(\hat{t}_y)$ will be highly efficient. On the other hand, if there is no correlation between y and $\hat{y}$, then no improvement will be achieved by using the calibrated weights $\boldsymbol{w}_c^{(k)}$; see also Theorem 3 in Section 4.

One of the drawbacks of the chi-square distance $\Phi(\boldsymbol{w}_c^{(k)}, \boldsymbol{w}_0^{(k)})$ in *Step* 4 is that some of the resulting calibrated weights could take negative values. To avoid negative weights, we propose replacing the chi-square distance in (3.7) by the following minimum entropy distance

$$D(\boldsymbol{w}_c^{(k)}, \boldsymbol{w}_0^{(k)}) = -\sum_{i\in\mathcal{S}_k} \tau_i^{-1} \left\{ w_{i0}^{(k)} \log\left(\frac{w_{ic}^{(k)}}{w_{i0}^{(k)}} \right) - w_{ic}^{(k)} + w_{i0}^{(k)} \right\} \tag{3.12}$$

for two reasons. First, the calibrated weights $w_{ic}^{(k)}$ are guaranteed to be positive. Second, there exists a well-behaved computational algorithm for this constrained minimization problem. It can be shown that $\boldsymbol{w}_c^{(k)}$ minimizing $D(\boldsymbol{w}_c^{(k)}, \boldsymbol{w}_0^{(k)})$ subject to (3.8) are given by

$$w_{ic}^{(k)} = \frac{w_{i0}^{(k)} / \tau_i}{1 + \boldsymbol{\lambda}'\boldsymbol{z}_i}, \tag{3.13}$$

where the Lagrange multiplier $\boldsymbol{\lambda}$ is the solution to

$$g(\boldsymbol{\lambda}) = \sum_{i\in\mathcal{S}} \frac{w_{i0}^{(k)}\boldsymbol{z}_i / \tau_i}{1 + \boldsymbol{\lambda}'\boldsymbol{z}_i} - \hat{t}_{\boldsymbol{z}}^{(k)} = \boldsymbol{0}. \tag{3.14}$$

An efficient computational algorithm for finding the solution $\boldsymbol{\lambda}$ to (3.14) can be found in Wu (2004) and a related R function can be obtained by a minor modification of the R function presented in Wu (2005).

4 Validity

In this section we provide some general discussion on the validity of the replication variance estimator. Let $\theta = f(t_y)$ be a finite population parameter, which is a smooth function of the population total $t_y = \sum_{i=1}^N y_i$. We assume that $\hat{\theta} = f(\hat{t}_y)$ is used to estimate θ, where $\hat{t}_y$ is the

Horvitz-Thompson estimator of t_y defined in (2.1). The replication variance estimator of $\hat{\theta}$ is constructed by

$$v_R(\hat{\theta}) = \sum_{k=1}^{L} c_k (\hat{\theta}^{(k)} - \hat{\theta})^2, \tag{4.1}$$

where $\hat{\theta}^{(k)} = f(\hat{t}_y^{(k)})$ and $\hat{t}_y^{(k)}$ is the k^{th} replicate of $\hat{t}_y$.

To explore the asymptotic properties of the replication variance estimator (4.1), we assume a sequence of the finite populations and the survey samples, as described in Isaki and Fuller (1982). The finite populations and the sampling designs satisfy following regularity conditions.

C1. For any population characteristics $\boldsymbol{u}_i$ with bounded second moments,

$$\sum_{i \in \mathcal{S}} w_i \boldsymbol{u}_i \boldsymbol{u}_i' - \sum_{i=1}^{N} \boldsymbol{u}_i \boldsymbol{u}_i' = O_p(n^{-1/2} N).$$

C2. The design weights are uniformly bounded. That is, $K_1 < N^{-1} n w_i < K_2$ for all i and any n, where K_1 and K_2 are fixed constants.

C3. $nV(N^{-1}\hat{t}_y)$ is bounded.

C4. For any y with bounded fourth moments, the replication variance estimator $v_R(\hat{t}_y) = \sum_{k=1}^{L} c_k (\hat{t}_y^{(k)} - \hat{t}_y)^2$ satisfies

$$E[\{c_k (\hat{t}_y^{(k)} - \hat{t}_y)^2\}^2] < KL^{-2}\{V(\hat{t}_y)\}^2 \tag{4.2}$$

for some K, uniformly in $k = 1, \ldots, L$,

$$\max_k c_k^{-1} = O(L), \tag{4.3}$$

and

$$E\left[\left\{\frac{v_R(\hat{t}_y)}{V(\hat{t}_y)} - 1\right\}^2\right] = o(1). \tag{4.4}$$

Condition (4.2) ensures that no single replicate dominate the others. Condition (4.3) controls the order of the factor c_k. Condition (4.4) implies that $v_R(\hat{t}_y)$ is a consistent estimator of $V(\hat{t}_y)$. Conditions (4.2) - (4.4) were also used in Kim, Navarro and Fuller (2006).

Using the above regularity conditions, the following theorem proves the consistency of the replication variance estimator in the form of (4.1).

Theorem 2. Let $\theta = f(t_y)$ be the parameter of interest and $\hat{\theta} = f(\hat{t}_y)$, where $f(\cdot)$ is a smooth function with derivative continuous at t_y. Under the regularity conditions described above, the variance estimator $v_R(\hat{\theta})$ in (4.1) satisfies

$$\frac{v_R(\hat{\theta})}{V(\hat{\theta})} = 1 + o_p(1). \tag{4.5}$$

Proof. See Appendix A.

We now prove the validity of the improved variance estimator $v_C(\hat{t}_y)$ proposed in Section 3.2. For simplicity, we assume that $v_1(\hat{t}_y)$ is a fully efficient estimator of the variance $V(\hat{t}_y)$ for $\hat{t}_y = \sum_{i \in \mathcal{S}} w_i y_i$. We also assume that $v_0(\hat{t}_y)$, defined in (3.4), satisfies

$$E^*\{v_0(\hat{t}_y)\} = v_1(\hat{t}_y), \tag{4.6}$$

where $E^*(\cdot)$ denotes expectation under the random selection of the L_0 replicates from the L sets of fully efficient replication weights, as discussed in Section 3.1. If $v_1(\hat{t}_y)$ is asymptotically unbiased, then $v_0(\hat{t}_y)$ is also asymptotically unbiased by (4.6). For the delete-a-group jackknife, condition (4.6) can be understood as $E\{v_0(\hat{t}_y)\} = E\{v_1(\hat{t}_y)\}$ and $V\{v_0(\hat{t}_y)\} \geq V\{v_1(\hat{t}_y)\}$.

Theorem 3. Assume that the initial variance estimator $v_0(\hat{t}_y)$ defined in (3.4) satisfies (4.6). Assume that the improved variance estimator $v_C(\hat{t}_y) = \sum_{k=1}^{L_0} c_{k0}(\hat{t}_{yc}^{(k)} - \hat{t}_y)^2$ is computed using the calibrated replication weights as described in Section 3.2, with the choice of τ_i satisfying $\text{Cov}(\hat{t}_e, \hat{t}_z) \doteq \mathbf{0}$. By ignoring smaller order terms, we have

$$E\{v_C(\hat{t}_y)\} = E\{v_1(\hat{t}_y)\} \tag{4.7}$$

and

$$V\{v_1(\hat{t}_y)\} \leq V\{v_C(\hat{t}_y)\} \leq V\{v_0(\hat{t}_y)\}. \tag{4.8}$$

Proof. See Appendix B.

For a general parameter $\theta = f(t_y)$, we let $\hat{\theta}_c^{(k)} = f(\hat{t}_{yc}^{(k)})$ and compute $v_C(\hat{\theta}) = \sum_{k=1}^{L_0} c_{k0}(\hat{\theta}_c^{(k)} - \hat{\theta})^2$. Validity of $v_C(\hat{\theta})$ can be established by combining results from Theorem 2 and Theorem 3.

5 Extension to some balanced sampling designs

We now consider sampling designs which are balanced in $\boldsymbol{x}_i$ in the sense that $\hat{t}_{\boldsymbol{x}} = \sum_{i\in\mathcal{S}} \boldsymbol{x}_i/\pi_i = t_{\boldsymbol{x}}$ holds exactly or nearly exactly, where $\boldsymbol{x}_i$ is a q-dimensional vector and $t_{\boldsymbol{x}} = \sum_{i\in\mathcal{U}} \boldsymbol{x}_i$ is known. We assume that the first element of $\boldsymbol{x}_i$ is equal to π_i, which implicitly assumes that the survey design has fixed sample size. Tillé (2006) provides a comprehensive account of balanced sampling designs.

Deville and Tillé (2005) argue that $\hat{t}_y = \sum_{i\in\mathcal{S}} y_i/\pi_i$ under balanced sampling has a variance that can be approximated by its variance under conditional Poisson sampling. Breidt and Chauvet (2011) using the same approximation derived

$$v(\hat{t}_y) = \frac{n}{n-q}\sum_{i\in\mathcal{S}}(1-\pi_i)\left(\frac{y_i}{\pi_i} - \frac{\tilde{y}_i}{\pi_i}\right)^2, \tag{5.1}$$

where $\tilde{y}_i = \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_P$ and $\hat{\boldsymbol{\beta}}_P = \left\{\sum_{i\in\mathcal{S}}(1-\pi_i)\pi_i^{-2}\boldsymbol{x}_i\boldsymbol{x}_i'\right\}^{-1}\sum_{i\in\mathcal{S}}(1-\pi_i)\pi_i^{-2}\boldsymbol{x}_i y_i$. Roughly speaking, the variance formula (5.1) can be interpreted as approximating $\hat{t}_y$ under the balanced sampling design by a generalized regression estimator under Poisson sampling. That is, $V(\hat{t}_y) \doteq V(\hat{t}_P)$, where $\hat{t}_P = \hat{t}_y + (t_{\boldsymbol{x}} - \hat{t}_{\boldsymbol{x}})'\hat{\boldsymbol{\beta}}_P$. For a formal justification on this approximation, see Fuller (2009b).

The variance formula (5.1) can be used to derive replication weights. To see this, we re-express (5.1) as a jackknife replication variance estimator

$$v_J(\hat{t}_y) = \sum_{k=1}^{n} c_k(\tilde{t}_y^{(k)} - \hat{t}_y)^2, \tag{5.2}$$

where $\tilde{t}_y^{(k)} = \hat{t}_y^{(k)} + (t_{\boldsymbol{x}} - \hat{t}_{\boldsymbol{x}}^{(k)})'\hat{\boldsymbol{\beta}}_P^{(k)}, (\hat{t}_{\boldsymbol{x}}^{(k)}, \hat{t}_y^{(k)}) = \sum_{i\in\mathcal{S}^{(k)}}\pi_i^{-1}(\boldsymbol{x}_i', y_i)$,

$$\hat{\boldsymbol{\beta}}_P^{(k)} = \left\{\sum_{i\in\mathcal{S}^{(k)}}(1-\pi_i)\pi_i^{-2}\boldsymbol{x}_i\boldsymbol{x}_i'\right\}^{-1}\sum_{i\in\mathcal{S}^{(k)}}(1-\pi_i)\pi_i^{-2}\boldsymbol{x}_i y_i,$$

$c_k = (1-\pi_k)n/(n-q)$, and $\mathcal{S}^{(k)} = \mathcal{S} \setminus \{k\}$. To show the asymptotic equivalence between (5.1) and (5.2), we first note that

$$\tilde{t}_y^{(k)} - \hat{t}_y = (\hat{t}_y^{(k)} - \hat{t}_y) + (t_{\boldsymbol{x}} - \hat{t}_{\boldsymbol{x}}^{(k)})'\hat{\boldsymbol{\beta}}_P + (t_{\boldsymbol{x}} - \hat{t}_{\boldsymbol{x}}^{(k)})'(\hat{\boldsymbol{\beta}}_P^{(k)} - \hat{\boldsymbol{\beta}}_P).$$

Under certain regularity conditions, we have $\hat{\boldsymbol{\beta}}_P^{(k)} = \hat{\boldsymbol{\beta}}_P + O_p(n^{-1})$ and $t_{\boldsymbol{x}} - \hat{t}_{\boldsymbol{x}}^{(k)} = \boldsymbol{x}_k/\pi_k = O_p(n^{-1}N)$. Here we used the condition $t_{\boldsymbol{x}} = \hat{t}_{\boldsymbol{x}}$ under the balanced sampling design. It follows that $\tilde{t}_y^{(k)} - \hat{t}_y = -\pi_k^{-1}(y_k - \boldsymbol{x}_k'\hat{\boldsymbol{\beta}}_P) + O_p(n^{-2}N)$, and $v_J(\hat{t}_y)$ in (5.2) is asymptotically equivalent to $\sum_{k=1}^n c_k\pi_k^{-2}(y_k - \tilde{y}_k)^2$, which equals $v(\hat{t}_y)$ given by (5.1). The variance formula (5.2) is quite useful because it makes the construction of the replication weights quite straightforward for balanced sampling designs. When n is large, the number of replicates can be reduced by using the weight-calibration method described in Section 3.2. Simulation results based on the rejective Poisson sampling of Fuller (2009b), not reported here to save space, showed that the proposed replication variance estimator performs very well.

6 Simulation study

In this section we report results from a simulation study. We consider a synthetic finite population of size $N = 2{,}248$ families using a real data set of Statistics Canada's 2000 Family Expenditure Survey for the province of Ontario. For the i^{th} selected family, the data set contains observations on several variables, including x_{i1}: the number of persons in the family; x_{i2}: the number of children (age < 15); x_{i3}: the number of youths (age 15 - 24); x_{i4}: the total annual income after taxes; y_{i1}: the total annual expenditure; y_{i2}: the annual expenditure on clothing; y_{i3}: the annual expenditure on furnishings and equipment.

We consider three population parameters for comparing different versions of replication variance estimators. The first is the population total of overall annual expenditures, *i.e.*, $\theta_1 = t_{y1} = \sum_{i=1}^N y_{i1}$. The second is the ratio of population totals of expenditures on clothing and on furnishings and equipment, *i.e.*, $\theta_2 = t_{y2}/t_{y3} = \left(\sum_{i=1}^N y_{i2}\right)\Big/\left(\sum_{i=1}^N y_{i3}\right)$. Note that $\theta_2 = \mu_{y2}/\mu_{y3}$. The third is the population correlation coefficient $\theta_3 = \rho(y_1, y_2)$ between the

overall annual expenditure (y_1) and the annual expenditure on clothing (y_2). For each parameter, several replication variance estimators were evaluated through simulation.

We investigate two approaches of replication variance estimation. For the first one, the initial L sets of the replication weights are constructed using the general method described in Section 2. For the second one, the $L = n$ sets of standard delete-1 jackknife replication weights are used to produce fully efficient variance estimators.

Case I. Unequal probability samples are selected by the Rao-Sampford PPS sampling method (Rao 1965; Sampford 1967), with inclusion probabilities π_i proportional to the total annual income x_{i4}. One of the attractive features of the Rao-Sampford method is that the second order inclusion probabilities π_{ij} can be computed exactly. The general procedure described in Section 2 is used to create $L = n$ sets of fully efficient replication weights, and the corresponding variance estimator is denoted as v_R. Those weights are used as the basis to compute and compare different versions of variance estimators $v_R^{(l)}, l = 1, 2, 3, 4$ described in Section 3.1, based on a smaller number L_0 sets of weights. We restrict L_0 to be 25 or 50.

The calibrated replication variance estimator described in Section 3.2 is denoted as v_C. The initial L_0 sets of weights are selected from the original L sets of weights by simple random sampling, and $\boldsymbol{z}_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4}, y_{i2}, y_{i3})'$ is used as calibration variables. Under this setting, the first parameter θ_1 is not directly related to $\boldsymbol{z}$ but the second parameter θ_2 is defined as a nonlinear but smooth function of $t_{\boldsymbol{z}}$. The third parameter $\rho(y_1, y_2)$ is more complex and involves population quantities not included in $t_{\boldsymbol{z}}$.

Case II. The population of $N = 2{,}248$ units is first duplicated 10 times, to create a larger population with 22,480 units. Simple random samples of $n = 100$, 200 or 400 are selected from the population. The sampling fractions are less than 2%. Under such scenarios the standard n sets of delete-1 jackknife weights provide fully efficient variance estimator v_J. Let $v_J^{(1)}$ be the variance estimator using L_0 sets of weights, randomly selected from the n sets of jackknife

weights. Let $v_J^{(C)}$ be the variance estimator using the L_0 sets of weights plus calibration over the $\boldsymbol{z}$ variables.

For each simulated sample of size n and a particular population parameter θ, we compute design-based estimator $\hat{\theta}$ and different versions of variance estimators. The process is repeated B times, independently, with $B = 5{,}000$ for *Case I* and $B = 10{,}000$ for *Case II*. The true variance $V = V(\hat{\theta})$ is approximated by $V \doteq B^{-1}\sum_{b=1}^{B}(\hat{\theta}_b - \theta)^2$, where $\hat{\theta}_b$ is calculated from the b^{th} simulated sample, using another independent B simulated samples. Simulation results show that the bias of $\hat{\theta}$ is negligible for all three parameters. Performances of a variance estimator v are measured by the simulated coverage probability of the 95% normal theory confidence interval, computed as $\text{CP} = B^{-1}\sum_{b=1}^{B} I[\hat{\theta}_b - 1.96(v_b)^{1/2} \le \theta \le \hat{\theta}_b + 1.96(v_b)^{1/2}]$, the average length of the interval $\text{AL} = B^{-1}\sum_{b=1}^{B} 2 \times 1.96(v_b)^{1/2}$, and the Relative Root Mean Square Error (RRMSE), computed as $\text{RRMSE} = \{\text{MSE}(v)\}^{1/2} / V$, where v_b is the variance estimator v computed from the b^{th} simulated sample, and $\text{MSE}(v) = B^{-1}\sum_{b=1}^{B}(v_b - V)^2$.

The simulated coverage probabilities are reported in Tables 6.1 and 6.2. The fully efficient variance estimator v_R and v_J provides good coverage for all scenarios except for $\rho(y_1, y_2)$ with *Case I* where the coverage is a bit low. The variance estimators $v_R^{(l)}, l = 1, 2, 3, 4$ and $v_J^{(1)}$ based on L_0 sets of weights seem to work for θ_1, to certain degree for θ_2 as well, but none is working for $\theta_3 = \rho(y_1, y_2)$. The calibrated estimator v_C provides satisfactory coverage for all scenarios for *Case I*. As for the calibrated estimator $v_J^{(C)}$ with *Case II*, it works very well for θ_1 and θ_2, but none are working well for $\theta_3 = \rho(y_1, y_2)$.

It should be noted that the definition of $\rho(y_1, y_2)$ involves population means over three derived variables y_1^2, y_2^2 and $y_1 y_2$. When those three variables are also included at the calibration stage, in addition to $\boldsymbol{z}$, the resulting variance estimator is denoted as $v_J^{(C+)}$ for *Case II*. It turns out that $v_J^{(C+)}$ provides much better results for θ_3 and also improved results for θ_1 and θ_2.

Also included in Tables 6.1 and 6.2 are the average length of the confidence intervals using $v_C, v_J^{(C)}$ and $v_J^{(C+)}$. The results (AL, in parentheses) are relative to the length of the interval using v_R (Table 6.1) or v_J (Table 6.2), with a value (say) 1.05 indicating 5% increase in length. It can be seen that the calibrated variance estimators produce confidence intervals which are either comparable in length to the intervals using $v_R (v_J)$ or slightly wider, depending on the parameter and/or sample sizes.

The relative root mean square errors (RRMSE) of variance estimators are presented in Tables 6.3 and 6.4. The results seem to depend not only on the parameter and its estimator but also the sampling design and the replication method. For *Case I*, the variance estimator v_C, which is of primary interest, is more stable than v_R for θ_1, almost the same for θ_2, and is less stable for θ_3. Because y_{i1} is well explained by $\boldsymbol{z}_i, v_C$ is quite efficient for estimating the variance of $\hat{\theta}_1 = \hat{t}_{y1}$. For *Case II*, $v_J^{(C)}$ and $v_J^{(C+)}$ are similar to each other but both are less stable than v_J.

Table 6.1
Coverage probabilities of 95% confidence intervals (*Case I*)

θ	L_0	n	v_R	$v_R^{(1)}$	$v_R^{(2)}$	$v_R^{(3)}$	$v_R^{(4)}$	v_C (AL)
t_{y1}	25	50	93.9	92.9	93.1	92.4	93.1	94.3 (1.03)
		100	94.4	92.0	92.4	91.9	93.0	93.4 (1.01)
		150	95.1	91.5	91.9	92.1	93.2	93.7 (0.99)
	50	100	94.5	93.2	93.2	93.4	93.6	94.1 (1.01)
		150	95.1	93.0	93.3	93.5	93.8	94.5 (0.99)
μ_{y2} / μ_{y3}	25	50	92.6	91.0	91.2	90.6	91.0	92.9 (1.02)
		100	93.7	91.1	91.2	89.6	90.8	93.7 (1.01)
		150	94.3	91.1	90.7	89.5	90.8	94.3 (1.00)
	50	100	93.6	92.6	92.5	91.9	92.5	93.8 (1.01)
		150	94.2	92.7	92.6	91.9	92.9	94.3 (1.00)
$\rho(y_1, y_2)$	25	50	89.0	85.7	85.6	79.3	81.9	91.3 (1.14)
		100	90.5	85.4	85.3	78.6	81.5	92.1 (1.17)
		150	90.7	84.6	84.5	76.9	81.6	91.9 (1.17)
	50	100	90.4	88.2	88.2	83.2	85.8	92.9 (1.18)
		150	90.7	87.5	87.6	81.7	84.8	93.4 (1.18)

v_R: The fully efficient replication variance estimator (Section 2); $v_R^{(l)}$, $l = 1, 2, 3, 4$: replication variance estimators based on L_0 sets of weights (Section 3.1); v_C: replication variance estimator based on L_0 sets of calibrated weights (Section 3.2); AL: average length of the confidence interval relative to the one using v_R.

Table 6.2
Coverage probabilities of 95% confidence intervals (*Case II*)

θ	L_0	n	v_J	$v_J^{(1)}$	$v_J^{(C)}$ (AL)	$v_J^{(C+)}$ (AL)
t_{y1}	25	100	94.4	92.0	94.4 (1.02)	94.9 (1.07)
		200	95.0	92.4	95.0 (1.01)	95.2 (1.03)
		400	95.3	92.5	95.1 (0.99)	95.3 (1.01)
	50	100	94.1	93.1	94.2 (1.02)	94.8 (1.07)
		200	94.7	93.3	94.8 (1.01)	95.0 (1.04)
		400	94.7	93.4	94.5 (0.99)	94.8 (1.02)
μ_{y2} / μ_{y3}	25	100	92.6	87.3	92.6 (1.05)	93.3 (1.11)
		200	93.6	86.8	93.3 (1.02)	93.7 (1.07)
		400	94.1	86.8	93.8 (0.99)	94.1 (1.04)
	50	100	92.8	90.3	92.9 (1.06)	94.2 (1.11)
		200	93.9	89.8	93.8 (1.03)	94.3 (1.08)
		400	94.1	89.6	93.8 (1.00)	94.1 (1.05)
$\rho(y_1, y_2)$	25	100	92.5	78.0	89.4 (1.06)	91.7 (1.09)
		200	92.7	72.3	86.3 (1.00)	91.4 (1.05)
		400	93.2	71.2	84.5 (0.95)	92.1 (1.04)
	50	100	92.2	84.5	92.2 (1.09)	92.5 (1.11)
		200	92.8	80.5	90.3 (1.05)	92.2 (1.08)
		400	93.1	77.4	88.1 (1.00)	92.6 (1.06)

v_J: The delete-1 jackknife variance estimator; $v_J^{(1)}$: replication variance estimator based on L_0 sets of jackknife weights; $v_J^{(C)}$: replication variance estimator based on L_0 sets of calibrated jackknife weights; $v_J^{(C+)}$: replication variance estimator based on L_0 sets of calibrated jackknife weights, with added variables for weight-calibration; AL: average length of the confidence interval relative to the one using v_J.

Table 6.3
Relative Root Mean Square Errors (RRMSE, *Case I*)

θ	L_0	n	v_R	$v_R^{(1)}$	$v_R^{(2)}$	$v_R^{(3)}$	$v_R^{(4)}$	v_C
t_{y1}	25	50	1.84	2.76	2.24	1.99	1.86	1.43
		100	1.32	2.34	1.67	1.89	1.40	0.83
		150	1.19	1.99	1.34	1.37	1.46	0.87
	50	100	1.32	1.91	1.69	1.63	1.35	0.92
		150	1.19	1.81	1.50	1.62	1.24	0.73
μ_{y2} / μ_{y3}	25	50	0.72	0.89	0.88	1.07	0.89	0.74
		100	0.45	0.78	0.77	0.99	0.72	0.46
		150	0.41	0.93	0.87	1.01	0.74	0.41
	50	100	0.46	0.60	0.60	0.77	0.56	0.46
		150	0.41	0.70	0.68	0.70	0.54	0.41
$\rho(y_1, y_2)$	25	50	0.65	0.79	0.83	1.45	1.26	0.96
		100	0.65	1.12	1.16	2.20	1.37	1.24
		150	0.59	1.29	1.34	2.27	1.43	1.50
	50	100	0.65	0.84	0.88	1.63	0.95	1.03
		150	0.59	0.88	0.94	1.48	1.05	1.12

v_R: The fully efficient replication variance estimator (Section 2); $v_R^{(l)}$, $l = 1, 2, 3, 4$: replication variance estimators based on L_0 sets of weights (Section 3.1); v_C: replication variance estimator based on L_0 sets of calibrated weights (Section 3.2).

Table 6.4
Relative Root Mean Square Errors (RRMSE, *Case II*)

θ	L_0	n	v_J	$v_J^{(1)}$	$v_J^{(C)}$	$v_J^{(C+)}$
t_{y1}	25	100	0.29	0.56	0.56	0.66
		200	0.21	0.57	0.47	0.53
		400	0.15	0.56	0.20	0.41
	50	100	0.29	0.41	0.50	0.57
		200	0.21	0.41	0.39	0.44
		400	0.15	0.41	0.17	0.35
μ_{y2} / μ_{y3}	25	100	0.78	1.58	1.90	1.98
		200	0.56	1.44	1.41	1.57
		400	0.39	1.54	0.87	1.40
	50	100	0.81	1.12	1.61	1.67
		200	0.57	1.10	1.22	1.32
		400	0.38	1.04	0.72	1.00
$\rho(y_1, y_2)$	25	100	1.02	2.43	2.71	2.72
		200	0.74	2.44	2.64	2.57
		400	0.47	2.51	2.52	2.42
	50	100	1.04	1.64	1.97	2.12
		200	0.71	1.75	2.01	1.96
		400	0.48	1.76	1.83	1.73

v_J: The delete-1 jackknife variance estimator; $v_J^{(1)}$: replication variance estimator based on L_0 sets of jackknife weights; $v_J^{(C)}$: replication variance estimator based on L_0 sets of calibrated jackknife weights; $v_J^{(C+)}$: replication variance estimator based on L_0 sets of calibrated jackknife weights, with added variables for weight-calibration.

7 Some concluding remarks

Replication methods offer an asymptotically equivalent alternative to linearization methods but are operationally more convenient and flexible. We focused on population parameters that are smooth functions of means or totals. Our theoretical results and limited simulation studies showed that the proposed strategies for constructing sparse and efficient replication weights work well for variance estimation and confidence intervals. Nevertheless, there are a number of issues which require further investigation. First, for complex parameters such as population correlation coefficients, sparse replication variance estimators are not very stable. Second, further evidences on the effectiveness of the proposed strategies for large complex surveys in conjunction to the use of general bootstrap or jackknife weights are needed. Third, it is not clear whether the sparse replication weights will be efficient for parameters that are not smooth functions of means or

totals, such as population quantiles, for which normal theory confidence intervals are known to be inefficient (Sitter and Wu 2001).

Another important issue is the potential application of the proposed methods for parameters and estimators defined through estimating equations. Let $\boldsymbol{\theta}$ be defined as the solution to

$$U_N(\boldsymbol{\theta}) = \sum_{i=1}^{N} u_i(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}) = \mathbf{0}. \tag{7.1}$$

Let $\hat{\boldsymbol{\theta}}$ be obtained by solving a sample-based version of (7.1) given by

$$U_n(\boldsymbol{\theta}) = \sum_{i \in \mathcal{S}} w_i u_i(y_i, \boldsymbol{x}_i; \boldsymbol{\theta}) = \mathbf{0}. \tag{7.2}$$

Regression or logistic regression analyses using complex survey data can both be viewed special cases of the general forms given by (7.1) and (7.2). The usual sandwich-type variance of $\hat{\boldsymbol{\theta}}$ is given by

$$V(\hat{\boldsymbol{\theta}}) \doteq \left\{ \frac{\partial U_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\}^{-1} V\{U_n(\boldsymbol{\theta})\} \left\{ \frac{\partial U_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\}^{-1} \tag{7.3}$$

A variance estimator $v(\hat{\boldsymbol{\theta}})$ can now be obtained if we substitute $\partial U_N(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ by $\partial U_n(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ and estimate $V\{U_n(\boldsymbol{\theta})\}$ by applying replication variance estimation method to $\hat{U}_n = \sum_{i \in \mathcal{S}} w_i \boldsymbol{u}_i$ with $\boldsymbol{u}_i = u_i(y_i, \boldsymbol{x}_i; \hat{\boldsymbol{\theta}})$. For detailed discussions on estimating equations and survey sampling, see, for instance, Binder (1983), Skinner (1989), and Godambe and Thompson (2009), among others.

Achieving efficient variance estimation using a limited number of sets of replication weights is an important research problem with both theoretical and practical significance. The fully efficient replication weights constructed using the procedure described in Section 2 can be treated as initial sets of weights if the sample size n is large. In principle, our proposed strategies in Section 3 for producing sparse and efficient replication weights can be combined with other initial sets of replication weights, including bootstrap weights (Shao 1996) or delete-a-group

jackknife (Kott 2001). One should also include as many relevant variables as possible in the calibration step, so that the final calibrated replication weights are not only sparse but also efficient in providing variance estimators for a large class of estimators. Extensions of the proposed methods to handle calibration weights or nonresponse adjustment are currently under investigation.

Acknowledgements

We thank two anonymous referees and the associate editor for their very helpful comments. This work started with initial discussions between the first author J.K. Kim and Professor Randy Sitter of Simon Fraser University who was tragically lost at sea during a kayak trip in 2007. The authors would like to dedicate this paper to the memory of Professor Sitter who was also the PhD thesis supervisor of the second author C. Wu. The research of J.K. Kim was partially supported by a Cooperative Agreement between the US Department of Agriculture Natural Resources Conservation Service and Iowa State University. The research of C. Wu was supported by grants from the Natural Sciences and Engineering Research Council of Canada and Mathematics of Information Technology and Complex Systems.

Appendix

A Proof of Theorem 2

By assumption (4.2), we have

$$\max_{1\le k\le L} c_k(\hat{t}_y^{(k)} - \hat{t}_y)^2 = O_p(L^{-1}n^{-1}N^2),$$

which, combined with (4.3), implies that

$$\max_{1\le k\le L}(\hat{\mu}_y^{(k)} - \hat{\mu}_y) = o_p(1), \tag{A.1}$$

where $\hat{\mu}_y^{(k)} = N^{-1}\hat{t}_y^{(k)}$ and $\hat{\mu}_y = N^{-1}\hat{t}_y$. Let $g(\mu_y) = f(N\mu_y)$. We can write

$$\hat{\theta}^{(k)} - \hat{\theta} = g(\hat{\mu}_y^{(k)}) - g(\hat{\mu}_y) = \dot{g}(\hat{\mu}_y)(\hat{\mu}_y^{(k)} - \hat{\mu}_y) + Q_{nk}(\hat{\mu}_y^{(k)} - \hat{\mu}_y),$$

where $\dot{g}(\mu) = \partial g(\mu) / \partial \mu, Q_{nk} = \dot{g}(\mu_k^*) - \dot{g}(\hat{\mu}_y)$, and μ_k^* is an inner point on the line segment between $\hat{\mu}^{(k)}$ and $\hat{\mu}$. By (A.1), we have

$$\max_{1 \le k \le L}(\mu_k^* - \hat{\mu}_y) = o_p(1). \tag{A.2}$$

Define

$$D_\delta = \left\{ \mu \left| \max_k \left\| \mu_k^* - \mu \right\| < \delta \text{ and } \max_k \left\| \dot{g}(\mu_k^*) - \dot{g}(\mu) \right\| > \epsilon \right. \right\}.$$

By construction, we have, for any $\epsilon > 0$ and $\delta > 0$,

$$P\left\{ \max_k \left\| \dot{g}(\mu_k^*) - \dot{g}(\hat{\mu}_y) \right\| > \epsilon \right\} \le P(\hat{\mu}_y \in D_\delta) + P\left(\max_k \left\| \mu_k^* - \hat{\mu}_y \right\| \ge \delta \right).$$

By the continuity of $\dot{g}(\mu)$ at $\mu = \mu_y$ and the fact that $\hat{\mu}_y = \mu_y + o_p(1)$, we have that, for any $\epsilon > 0$, there exists a $\delta = \delta(\epsilon) > 0$ such that $P(\hat{\mu}_y \in D_\delta) = o(1)$. This, together with (A.2), implies that

$$\max_k \left\| \dot{g}(\mu_k^*) - \dot{g}(\hat{\mu}_y) \right\| = o_p(1). \tag{A.3}$$

Now, we have

$$\sum_{k=1}^{L} c_k (\hat{\theta}^{(k)} - \hat{\theta})^2 = A_n + B_n + 2C_n, \tag{A.4}$$

where

$$A_n = \sum_{k=1}^{L} c_k \{\dot{g}(\hat{\mu}_y)(\hat{\mu}_y^{(k)} - \hat{\mu}_y)\}^2,$$
$$B_n = \sum_{k=1}^{L} c_k \{Q_{nk}(\hat{\mu}_y^{(k)} - \hat{\mu}_y)\}^2, \text{ and}$$
$$C_n = \sum_{k=1}^{L} c_k \dot{g}(\hat{\mu}_y)(\hat{\mu}_y^{(k)} - \hat{\mu}_y)^2 Q_{nk}.$$

Note that (4.4) implies

$$\sum_{k=1}^{L} c_k (\hat{\mu}_y^{(k)} - \hat{\mu}_y)^2 / V(\hat{\mu}_y) = 1 + o_p(1). \tag{A.5}$$

By standard linearization arguments, we have $A_n / V(\hat{\theta}) \to 1$ in probability. Furthermore, by (A.3) and (A.5), we have $B_n / V(\hat{\theta}) = o_p(1)$ and $C_n / V(\hat{\theta}) = o_p(1)$. This establishes (4.5).

B Proof of Theorem 3

Combining (3.10) and (3.11) and ignoring terms of smaller order, we have

$$v_0(\hat{t}_y) - v_C(\hat{t}_y) \doteq \hat{\boldsymbol{\beta}}' v_0(\hat{t}_{\mathbf{z}})\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}' v_1(\hat{t}_{\mathbf{z}})\hat{\boldsymbol{\beta}} \doteq \boldsymbol{\beta}' v_0(\hat{t}_{\mathbf{z}})\boldsymbol{\beta} - \boldsymbol{\beta}' v_1(\hat{t}_{\mathbf{z}})\boldsymbol{\beta}.$$

where $\boldsymbol{\beta}$ is the probability limit of $\hat{\boldsymbol{\beta}}$. By (4.6), we have

$$E^*\{v_0(\hat{t}_{\mathbf{z}})\} = v_1(\hat{t}_{\mathbf{z}}), \tag{B.1}$$

where $E^*(\cdot)$ denotes expectation under random selection of the L_0 sets of weights conditional on the L sets of weights. Similarly, by (3.11), we have

$$v_1(\hat{t}_y) - v_C(\hat{t}_y) \doteq v_1(\hat{t}_e) - v_0(\hat{t}_e).$$

By (4.6) again, we have

$$E^*\{v_0(\hat{t}_e)\} = v_1(\hat{t}_e). \tag{B.2}$$

Let $\hat{d}_1 = v_C(\hat{t}_y) - v_1(\hat{t}_y)$, we have $E(\hat{d}_1) = 0$ by (B.2), which proves (4.7). Furthermore, by (B.2) again, we have $\mathrm{Cov}\{\hat{d}_1, v_1(\hat{t}_y)\} = 0$. Thus, we have

$$V\{v_C(\hat{t}_y)\} = V\{v_1(\hat{t}_y)\} + V(\hat{d}_1) \geq V\{v_1(\hat{t}_y)\}. \tag{B.3}$$

Similarly, we can also prove that $V\{v_0(\hat{t}_y)\} \geq V\{v_C(\hat{t}_y)\}$.

References

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

Breidt, F.J., and Chauvet, G. (2011). Improved variance estimation for balanced samples drawn via the cube method. *Journal of Statistical Planning and Inference*, 141, 411-425.

Campbell, C. (1980). A different view of the finite population estimation. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 319-324.

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology*, 25, 2, 193-203.

Deville, J.-C., and Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 128, 411-425.

Dippo, C.S., Fay, R.E. and Morganstein, D.H. (1984). Computing variances from complex samples with replicate weights. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Washington, DC, 489-494.

Fay, R.E. (1984). Some properties of estimators of variance based on replication methods. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Washington, DC, 495-500.

Fay, R.E., and Dippo, C.S. (1989). Theory and application of replicate weighting for variance calculations. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Washington, DC, 212-217.

Fuller, W.A. (1998). Replication variance estimation for two phase samples. *Statistica Sinica*, 8, 1153-1164.

Fuller, W.A. (2009a). *Sampling Statistics*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Fuller, W.A. (2009b). Some design properties of a rejective sampling procedure. *Biometrika*, 96, 933-944.

Godambe, V.P., and Thompson, M.E. (2009). Estimating functions and survey sampling. In *Handbook of Statistics*, (Eds., D. Pfeffermann and C.R. Rao), Sample Surveys: Inference and Analysis, North Holland, Vol. 29B, 83-101.

Gross, S. (1980). Median estimation in sample surveys. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, Washington, DC, 181-184.

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

Jang, D., and Eltinge, J.L. (2009). Use of within-primary-sample-unit variances to assess the stability of a standard design-based variance estimator. *Survey Methodology*, 35, 2, 235-245.

Kim, J.K., Navarro, A. and Fuller, W.A. (2006). Replication variance estimation for two-phase stratified sampling. *Journal of the American Statistical Association*, 101, 312-320.

Kott, P.S. (2001). The delete-a-group jackknife. *Journal of Official Statistics*, 17, 521-526.

Krewski, D., and Rao, J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *Annals of Statistics*, 9, 1010-1019.

Lu, W.W., Brick, J.M. and Sitter, R.R. (2006). Algorithms for constructing combining strata variance estimators. *Journal of the American Statistical Association*, 101, 1680-1692.

Lu, W.W., and Sitter, R.R. (2008). Disclosure risk and replication-based variance estimation. *Statistica Sinica*, 18, 1669-1687.

McCarthy, P.J., and Snowden, C.B. (1985). *The Bootstrap and Finite Population Sampling*. Vital and Health Statistics, Ser. 2, No. 95, Public Health Service Publication 85-1369, U.S. Government Printing Office, Washington, DC.

Preston, J. (2009). Rescaled bootstrap for stratified multistage sampling. *Survey Methodology*, 35, 2, 227-234.

Rao, J.N.K. (1965). On two simple schemes of unequal probability sampling without replacement. *Journal of the Indian Statistical Association*, 3, 173-180.

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.

Rust, K.F., and Kalton, G. (1987). Strategies for collapsing strata for variance estimation. *Journal of Official Statistics*, 3, 69-81.

Rust, K.F., and Rao, J.N.K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, 5, 283-310.

Sampford, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54, 499-513.

Shao, J. (1996). Resampling methods in sample surveys (with discussion). *Statistics*, 27, 203-254.

Shao, J. (2003). Impact of the bootstrap on sample surveys. *Statistical Science*, 18, 191-198.

Shao, J., and Tu, D. (1995). *The Jackknife and Bootstrap*. New York: Springer.

Sitter, R.R. (1992). A resampling procedure for complex survey data. *Journal of the American Statistical Association*, 87, 755-765.

Sitter, R.R., and Wu, C. (2001). A note on Woodruff confidence intervals for quantiles. *Statistics and Probability Letters*, 52, 353-358.

Skinner, C.J. (1989). Domain means, regression and multivariate analysis. In *Analysis of Complex Surveys*, (Eds., C.J. Skinner, D. Holt and T.M. Smith), New York: John Wiley & Sons, Inc., 59-88.

Tillé, Y. (2006). *Sampling Algorithms*. Springer Science + Business Media, Inc.

Wolter, K.M. (2007). *Introduction to Variance Estimation* (2nd Edition). New York: Springer-Verlag.

Wu, C. (2004). Some algorithmic aspects of the empirical likelihood method in survey sampling. *Statistica Sinica*, 14, 1057-1067.

Wu, C. (2005). Algorithms and R codes for the pseudo empirical likelihood method in survey sampling. *Survey Methodology*, 31, 2, 239-243.

Survey Methodology, June 2013
Vol. 39, No. 1, pp. 121-148
Statistics Canada, Catalogue No. 12-001-X

Estimation of the variance of cross-sectional indicators for the SILC survey in Switzerland

Anne Massiani[1]

Abstract

SILC (*Statistics on Income and Living Conditions*) is an annual European survey that measures the population's income distribution, poverty and living conditions. It has been conducted in Switzerland since 2007, based on a four-panel rotation scheme that yields both cross-sectional and longitudinal estimates. This article examines the problem of estimating the variance of the cross-sectional poverty and social exclusion indicators selected by Eurostat. Our calculations take into account the non-linearity of the estimators, total non-response at different survey stages, indirect sampling and calibration. We adapt the method proposed by Lavallée (2002) for estimating variance in cases of non-response after weight sharing, and we obtain a variance estimator that is asymptotically unbiased and very easy to program.

Key Words: SILC survey; Rotating panel; Inequality indices; Variance estimation; Weight-share method.

1 Introduction

SILC (Statistics on Income and Living Conditions) is an annual European survey designed to obtain indicators that are comparable from one country to another on poverty, social exclusion and living conditions within the population. For a detailed description of this survey, see Clemenceau and Museux (2006). In accordance with the recommendations of Eurostat (Eurostat 2003), the survey is conducted in Switzerland based on a four-panel rotation scheme; the first panel was surveyed in 2007. Each panel lasts four years, and every year one panel is replaced. When complete, a sample selected for a panel consists of approximately 3,600 households. This article will focus on cross-sectional estimation, for the population present in a year t, of the indicators selected by Eurostat on poverty and living conditions, such as the at-risk-of-poverty rate and the quintile share ratio. See Osier (2009) for a description of these indicators and Ardilly and Lavallée (2007) for a description of the cross-sectional approach in the context of the SILC survey. Under this

1. Anne Massiani, Institute of Statistics, University of Neuchâtel, Pierre-à-Mazel 7, 2000 Neuchâtel, Switzerland & Federal Office of Statistics, Espace de l'Europe 10, 2010 Neuchâtel, Switzerland. E-mail: anne.massiani@unine.ch.

cross-sectional approach, the change over time in the composition of the households in the panels selected is taken into account using the indirect sampling theory developed by Lavallée (2002).

In this study, the formulas presented for estimating the variance of indicators take account of the great complexity of SILC-Switzerland. The results obtained, while developed specifically for the survey conducted in Switzerland, are likely to interest the other participating countries since they use similar methods. The main factors taken into account in estimating the variance of the indicators are the non-linearity of the estimators, total non-response at different survey stages, indirect sampling and calibration. One solution for estimating the variance of non-linear indicators is to use linearization techniques (see Deville 1999). Formulas specifically adapted to the indicators selected by Eurostat have been developed by Osier (2009). An alternative formula for the quintile share ratio is available in Langel and Tillé (2011). Once linearization techniques are applied, there is still the problem of estimating the variance of a total in a complex survey design. One difficulty is the presence of non-response after weight sharing. Lavallée (2002) proposes an estimator of the variance of a total that takes this into account. However, this estimator is generally not unbiased, even when the response probabilities are known. In this study, we propose an adaptation that corrects this bias.

Section 2 briefly describes how the survey is conducted and its sampling design. Section 3 describes the weighting used. Section 4 describes how linearization techniques are applied to these estimators to obtain an approximation of their variance. Section 5 is devoted to the problem of estimating the variance of a total where there is non-response after weight sharing. The final formula used for variance estimation is given in Section 6. Finally, a numerical application is provided in Section 7, followed by the conclusions of this study.

2 Sample design and survey procedure

The survey is conducted in Switzerland using a four-panel rotation scheme. Each year t, a new sample of households $s_m^{A_1,t}$ is selected to replace an outgoing panel. The subscript m is used here

for samples of households; the subscript p will be reserved for samples of persons. In accordance with the notation of Lavallée (2002), the superscripts A_i refer to samples selected directly, while the superscript B will be used in the rest of the document for samples selected indirectly. The new panel selected will be followed for four years according to the following scenario:

- In the first year t, we first approach the households of $s_m^{A_1,t}$ to complete a preliminary questionnaire called the "grid". The variables identified in the grid are common to the entire household and relate primarily to its composition. The subsample of $s_m^{A_1,t}$ that responds to the grid is denoted $s_m^{A_2,t}$. The sample of individuals belonging to the households of $s_m^{A_2,t}$ is denoted $s_p^{A_2,t}$ and these individuals are called longitudinals. The households of $s_m^{A_2,t}$ reached in the first survey year t are then asked to complete a joint questionnaire for the entire household, called the household questionnaire.
- In the following three years $t + i$, for $i = 1$ to 3, an attempt is made to recontact the longitudinals of $s_p^{A_2,t}$ aged 16 and over to survey their households based on the household composition in year $t + i$, since households may change over time. All individuals in these households are integrated into the survey for year $t + i$ and are called cross-sectional individuals. Households for year $t + i$ that are reached by means of longitudinal – that is, households reached indirectly – are asked to again complete the grid and then the household questionnaire. However, it is not possible to recontact all the longitudinals, especially because some of them have moved, and this is a major cause of non-response and sample attrition.

We now set a given survey year t and adopt a cross-sectional approach, meaning that we are interested in the estimates that can be produced for the population present in year t. The sample surveyed during survey year t consists of four panels contacted for the first, second, third or fourth time respectively. Let $s_m^{B,\tau}$ denote the sample of households responding to the household questionnaire in the τ^{th} wave, for $\tau = 1, \ldots, 4$. As seen in Figure 2.1, the sample $s_m^{B,\tau}$ was contacted via the longitudinals selected in year $t_\tau = t - \tau + 1$. The shaded parts on the right represent the samples participating in the survey in year t (indirect sampling), while the samples on the left contain the initial households of the longitudinals through whom they were contacted (direct sampling).

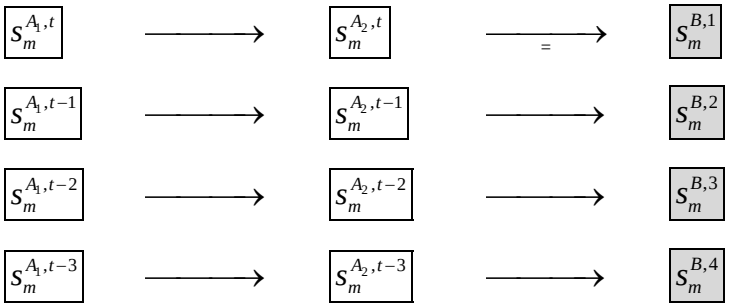


Figure 2.1 Panels comprising the sample surveyed in year t

Below is a more detailed description of the sampling design for $s_m^{A_2,t_\tau}$, for $\tau = 1, \ldots, 4$. Households' composition may change over time, and therefore in the rest of the document, we will use the notation k_1 to designate households in the first survey year t_τ and the notation k to designate those in survey year t so as to distinguish them. Each sample $s_m^{A_2,t_\tau}$ is obtained from two selection stages.

- **Stage 1:** a sample of households $s_m^{A_1,t_\tau}$ is selected according to a design stratified by major region, of which there are seven in Switzerland. Within each stratum, draws are conducted according to a simple design. Here, $\pi_{k_1}^{A_1}$ denotes the first order of probability of inclusion of household k_1 and $\pi_{k_1k_1'}^{A_1}$ denotes the second order of probability of inclusion of households k_1 and k_1'.
- **Stage 2:** the second selection stage is based on non-response to the grid in the first survey year t_τ. This non-response is modeled using a Poisson design on households, and we note for any household k_1:

$$\mathrm{P}(k_1 \in s_m^{A_2,t_\tau} \mid k_1 \in s_m^{A_1,t_\tau}) = q_{k_1}^a. \tag{2.1}$$

For all households k_1 and k_1', we define as follows:

$$\pi_{k_1}^{A_2} = \mathrm{P}(k_1 \in s_m^{A_2,t_\tau}) = \pi_{k_1}^{A_1} q_{k_1}^a \tag{2.2}$$

$$\pi_{k_1k_1'}^{A_2} = \mathrm{P}(k_1, k_1' \in s_m^{A_2,t_\tau}) = \begin{cases} \pi_{k_1}^{A_1} q_{k_1}^a & \text{if } k_1 = k_1' \\ \pi_{k_1k_1'}^{A_1} q_{k_1}^a q_{k_1'}^a & \text{if } k_1 \neq k_1'. \end{cases} \tag{2.3}$$

Finally, for all longitudinals j and j' belonging respectively to households k_1 and k_1', let

$$\pi_j^{A_2} = \mathrm{P}(j \in s_p^{A_2,t_\tau}) = \pi_{k_1}^{A_2} \quad \pi_{jj'}^{A_2} = \mathrm{P}(j, j' \in s_p^{A_2,t_\tau}) = \pi_{k_1 k_1'}^{A_2}. \tag{2.4}$$

For practical reasons, samples $s_m^{A_1,t_\tau}$, for $\tau = 1, \ldots, 4$ are drawn in such a way as to be disjoint, which means that this is also the case with samples $s_m^{A_2,t_\tau}$. On the other hand, the four $s_m^{B,\tau}$ samples are theoretically not necessarily disjoint, since there is a possibility that two longitudinals selected in two different waves will belong to the same household in the survey year. Considering the small sample sizes, this is extremely unlikely and has actually never been observed thus far. The fact is that when complete, a panel consists of approximately 3,600 households drawn with probabilities of selection $\pi_{k_1}^{A_1}$ for which the highest value is 0.0011. However, it should be noted that the methodology used would lend itself to processing non-disjoint $s_m^{B,\tau}$ samples. Let

$$s_m^B = \bigcup_{\tau=1}^{4} s_m^{B,\tau}$$

and a weight w_k, called the household cross-sectional weight, is calculated for each household k of s_m^B. The individuals in the households of s_m^B are denoted by s_p^B and each individual j of a household k of s_m^B receives the weight $w_j = w_k$.

The information used to calculate the poverty and social exclusion indexes is based on the income of each household k of s_m^B. Some components of the household's income are taken from the household questionnaire completed by all households of s_m^B. Other income components, such as wages, must again be obtained from questionnaires, this time from the individual questionnaires administered to members of the households of s_m^B. In the case of complete or partial non-response to individual questionnaires or of partial non-response to the household questionnaire, some components of the income of s_m^B households are missing. In some cases, especially where there is a lack of information on wages, missing values can be obtained from other sources based on administrative data. In the remaining cases, they are imputed. In this article, we do not take the effect of imputations into account. Under these conditions, the income of each household k of s_m^B is

known and the non-responses observed in the different stages of the survey are household total non-responses: non-response to the successive grids and total non-response to the household questionnaire. Since the poverty and social exclusion indexes are calculated for individuals and not households, the income of each household is divided by a factor that depends on the number of persons in the household and their age. The result of this calculation, called equivalent income, is assigned to each individual j of s_p^B. The poverty and social exclusion indexes are calculated on the basis of equivalent income and the weights w_j.

3 Cross-sectional household weighting

The weighting was carried out by Graf (2008). It utilizes non-response and calibration techniques (*cf.* Deville and Särndal 1992 for more information on calibration). It also uses the weight-share method (Lavallée 2002) to assign a weight to individuals reached indirectly. This weighting is carried out in five major steps. The objective of the four first step is to calculate a weight for the households in each panel $s_m^{B,\tau}$, for $\tau = 1, \ldots, 4$. These weights are calculated separately for each panel. The last step then serves to combine these panels. Below, we review the basic elements of each step for the situation in which the response probabilities for the different stages of the survey are known. Explanations on how these parameters are estimated are provided in the numerical application in Section 7.

1. *Calculate initial weight*
   Let $\tau \in \{1, \ldots, 4\}$ be fixed. For any longitudinal $j \in s_p^{A_2, t_\tau}$, we calculate the weight $w_j^{A_2}$ which takes account of non-response to the grid in wave 1:

$$w_j^{A_2} = \frac{1}{\pi_j^{A_2}}. \tag{3.1}$$

2. *Adjust for non-response to the grid in the survey year*
   Sample attrition is observed between the first wave and the survey year. First, it is not possible to recontact all the longitudinals of $s_p^{A_2, t_\tau}$. Second, not all the households that we manage to recontact via the longitudinals agree to fill out the grid again. We

designate as q_k^b the probability of household k responding to the grid in the survey year, conditional on $s_p^{A_2,t_\tau}$. For the panel responding in the first wave, the q_k^b values are all equal to 1, since there has not yet been any attrition. For any longitudinal j belonging to household k, let $q_j^b = q_k^b$. We then calculate the weight:

$$w_j^g = w_j^{A_2} \frac{1}{q_j^b}. \tag{3.2}$$

For reasons that will be explained in Section 7, the estimation of response probabilities q_j^b is a delicate problem. Because of variations in the information available, these estimates are not all equal for the longitudinals in a given household k, and this does not accord with the response mechanism. This difficulty, and how it is dealt with in estimating the variances of the indicators, will be discussed in Section 7.

3. *Weight sharing*

Following the methodology of Lavallée (2002, Chapter 6), we introduce the concept of the initially present cohabitant: this is an individual who was not selected in wave 1, but who was included in the target population during the selection of wave 1. Newborns and immigrants are called initially absent cohabitants. Let L_k be the number of longitudinals and P_k be the number of cohabitants initially present in a household k in the survey year. The weight of a household k responding to the grid in the survey year is calculated using weight sharing as follows:

$$w_k^p = \frac{1}{L_k + P_k} \sum_{j=1}^{L_k} w_j^g. \tag{3.3}$$

4. *Adjust for non-response to the household questionnaire*

The non-response observed between the grid for the survey year and the household questionnaire is modeled using a Poisson design on households. Let q_k^c be the probability that household k will respond to the household questionnaire, where we know that it responded to the grid for the survey year. For any household $k \in s_m^{B,\tau}$, the weight is then computed as follows:

$$w_k^{nr} = w_k^p \frac{1}{q_k^c}. \tag{3.4}$$

5. *Combine panels, then calibrate*

- Combine panels

  We want to obtain a weighting suitable for the amalgamation s_m^B of the four samples $s_m^{B,\tau}$, for $\tau = 1, \ldots, 4$. Accordingly, the weight w_k^{nr} computed in the

previous step must be divided, by order of magnitude, by a factor of 4. We have, for any household $k \in s_m^B$:

$$w_k^u = \phi_k^\tau \cdot w_k^{nr}, \tag{3.5}$$

where ϕ_k^τ values are allocation factors close to $1/4$ that must be optimized. Merkouris (2001) used variance minimization criteria to calculate the optimal values of ϕ_k^τ. These values depend on the variances, and hence on the design effects, associated with each panel. Since these design effects are unknown, we assume that within each major region, they are identical for each of the panels. This leads us to compute, within each major region, factors ϕ_k^τ that are proportional to the size of the sample of households responding in the τ^{th} wave.

- *Calibrate*
  Weights w_k^u are then calibrated on known margins on the population of households for the survey year. Let w_k be the calibrated weight thus obtained for household k.
- *Assign weights to individuals*
  For any individual j belonging to a household k of s_m^B, let

$$w_j = w_k.$$

Note 1: The different weighting steps have been presented in the order in which they are actually performed. Below we describe another way to obtain exactly the same final weights when the response probabilities q_j^b are known. In this case, since $q_j^b = q_k^b$ for all the longitudinals j in a given household k for the survey year, it is easy to see that the weight w_k^{nr} given by formula (3.4) can be rewritten as follows:

$$w_k^{nr} = \frac{1}{L_k + P_k}\left(\sum_{j=1}^{L_k} w_j^{A_2}\frac{1}{q_j^b}\right)\frac{1}{q_k^c} = w_k^{p^*}\frac{1}{q_k^b q_k^c}, \tag{3.6}$$

where

$$w_k^{p^*} = \frac{1}{L_k + P_k}\sum_{j=1}^{L_k} w_j^{A_2}. \tag{3.7}$$

Formula (3.6) shows that one can obtain the same weight w_k^{nr} as Graf (2008), and therefore the same final weights, by proceeding as follows:

- assign to household k for the survey year the shared weight $w_k^{p^*}$,

- in a single step, adjust for non-response to the grid and to the household questionnaire for the survey year by modeling it by a Poisson design on households of parameter $q_k^b q_k^c$, and applying the adjustment factor $1 / (q_k^b q_k^c)$.

This finding will be useful for the variance calculations in Section 5.

4 Linearization and approximation of variance

We want to estimate the variance $\text{var}(\hat{\Theta})$ of an estimator $\hat{\Theta}$ calculated on the sample of cross-sectional individuals s_p^B with assigned weights w_j. Lavallée (2002, pages 122-123) developed an asymptotic framework for a population surveyed indirectly. This framework lends itself to the use of linearization techniques (*cf.* Deville 1999) to obtain an approximation of the variance of a complex estimator calculated on a population surveyed indirectly. If $\hat{\Theta}$ is the estimator of one of the inequality indexes selected by Eurostat, linearization techniques are used to make estimation of the variance of $\hat{\Theta}$ equivalent to estimation of the variance of a total. The macros of Osier (2009) can be used for this purpose. Osier's linearization formulas are reviewed in Appendix A with respect to the four indicators considered in the numerical application in Section 7. Let ℓ_j denote the linearized values of $\hat{\Theta}$. We then have:

$$\text{var}(\hat{\Theta}) \simeq \text{var}\left(\sum_{j \in s_p^B} w_j \ell_j \right). \tag{4.1}$$

By using the residuals of the regression of the variable of interest in relation to the calibration variables, we can take account of the calibration effect in the calculations of variance (*cf.* Deville and Särndal 1992). Since the calibration variables x_k are defined at the household level, we first calculate the following for any household k for the survey year:

$$\ell_k = \sum_{j \in m_k} \ell_j,$$

where m_k designates all the members of household k, then we define $e_k = \ell_k - x_k^T \beta$, with the parameter β being calculated here based on all the households present in the population. We then have:

$$\mathrm{var}(\hat{\Theta}) \simeq \mathrm{var}\left(\sum_{j \in s_p^B} w_j \ell_j \right) = \mathrm{var}\left(\sum_{k \in s_m^B} w_k \ell_k \right) \simeq \mathrm{var}\left(\sum_{k \in s_m^B} w_k^u e_k \right). \tag{4.2}$$

Note 2: The linearized values ℓ_j introduce quantities that are calculated for the entire population, as may be seen, for example, in formula (A.6) in Appendix A. In accordance with the usual practice, the linearized values ℓ_j will ultimately be replaced by estimates $\hat{\ell}_j$. Similarly, since the quantities e_k are unknown, they will be replaced by estimates

$$\hat{e}_k = \hat{\ell}_k - x_k^T \hat{\beta}, \tag{4.3}$$

where

$$\hat{\ell}_k = \sum_{j \in m_k} \hat{\ell}_j$$

and

$$\hat{\beta} = \left(\sum_{k \in s_m^B} w_k x_k x_k^T \right)^{-1} \left(\sum_{k \in s_m^B} w_k x_k \hat{\ell}_k \right).$$

Finally, since the four samples $s_m^{B,\tau}$, for $\tau = 1, \ldots, 4$, which comprise s_m^B are reached through disjoint samples $s_m^{A_1,t_\tau}$, they are not strictly independent. However, we make the approximation that these four samples are independent, since the probabilities of selection $\pi_k^{A_1}$ are very low. We also assume that the allocation factors ϕ_k^τ that appear in formula (3.5) are not random. If we assume, for any household k:

$$e'_k = \phi_k^\tau \cdot e_k \tag{4.4}$$

and go back to (3.5), we can rewrite the amount that appears in the last member of (4.2) in the following form:

$$\sum_{k \in s_m^B} w_k^u e_k = \sum_{\tau=1}^{4} \sum_{k \in s_m^{B,\tau}} w_k^{nr} (\phi_k^\tau e_k) = \sum_{\tau=1}^{4} \left(\sum_{k \in s_m^{B,\tau}} w_k^{nr} e'_k \right). \tag{4.5}$$

This enables us to obtain the following approximation of the variance:

$$\text{var}(\hat{\Theta}) \simeq \sum_{\tau=1}^{4} \text{var}(\hat{T}_\tau), \tag{4.6}$$

where

$$\hat{T}_\tau = \sum_{k \in s_m^{B,\tau}} w_k^{nr} e'_k. \tag{4.7}$$

The four components of variance that appear in formula (4.6) can be computed and estimated in the same way. In the next section, we give an estimator of the variance of $\hat{T}_\tau$, for any τ.

5 Variance estimation and weight sharing

The calculations presented in this section are adaptations of the techniques developed by Lavallée (2002, Chapter 8.5) for the treatment of cluster non-response (CNR) in the context of indirect sampling. The results are given in the fictitious case where the probabilities of response at the different stages of the survey are known. The quantities e_k as well as the quantities e'_k defined by (4.4) are also assumed to be known. All these quantities will be replaced by estimates in Section 7. Let $\mathbf{1}_{\{k \in s_m^{B,\tau}\}}$ denote the indicator variable that is equal to 1 if household k present in survey year t is included in sample $s_m^{B,\tau}$ responding in the τ^{th} wave. Conditional upon the fact that an attempt was made to contact it via the longitudinals that it contains, household k belongs to $s_m^{B,\tau}$ if it responded to the grid and then to the questionnaire in survey year t. Therefore, we have:

$$E(\mathbf{1}_{\{k \in s_m^{B,\tau}\}} \mid s_p^{A_2,t_\tau}) = \text{P}(k \in s_m^{B,\tau} \mid s_p^{A_2,t_\tau}) = q_k^b q_k^c. \tag{5.1}$$

Using theorem 8.1 of Lavallée (2002, page 151) and Note 1, we can easily verify that the estimator (4.7) can be rewritten in the following form:

$$\hat{T}_\tau = \sum_{j \in s_p^{A_2,t_\tau}} w_j^{A_2} \hat{Z}_j = \sum_{j \in s_p^{A_2,t_\tau}} \frac{1}{\pi_j^{A_2}} \hat{Z}_j, \tag{5.2}$$

where we have noted, for any longitudinal $j \in s_p^{A_2,t_\tau}$ belonging to household k in the survey year:

$$\hat{Z}_j = \frac{e'_k}{L_k + P_k} \frac{1}{q_k^b} \frac{1}{q_k^c} \mathbf{1}_{\{k \in s_m^{B,\tau}\}}. \tag{5.3}$$

The estimator $\hat{T}_\tau$ is consequently reduced to a sum over individuals selected directly. We now decompose the variance of $\hat{T}_\tau$ in a standard way by conditioning on $s_p^{A_2,t_\tau}$:

$$\text{var}(\hat{T}_\tau) = \text{var}_{s_p^{A_2,t_\tau}} \left[\text{E}(\hat{T}_\tau \mid s_p^{A_2,t_\tau}) \right] + \text{E}_{s_p^{A_2,t_\tau}} \left[\text{var}(\hat{T}_\tau \mid s_p^{A_2,t_\tau}) \right]. \tag{5.4}$$

For any individual j who is present in the population during the year in which $s_p^{A_2,t_\tau}$ is drawn and who belongs to household k during the survey year, let

$$Z_j = \frac{e'_k}{L_k + P_k}. \tag{5.5}$$

Using (5.1), we verify that for all longitudinal j and j' included in $s_p^{A_2,t_\tau}$ and belonging to households k and k' respectively during the survey year, we have:

$$\text{E}(\hat{Z}_j \mid s_p^{A_2,t_\tau}) = Z_j \tag{5.6}$$

and

$$\text{cov}_{jj'} = \text{cov}\left[(\hat{Z}_j, \hat{Z}_{j'}) \mid s_p^{A_2,t_\tau} \right] = \begin{cases} 0 & \text{if } k \neq k' \\ \left(\dfrac{e'_k}{L_k + P_k} \right)^2 \dfrac{1 - q_k^b q_k^c}{q_k^b q_k^c} & \text{if } k = k'. \end{cases} \tag{5.7}$$

Formula (5.4) then becomes:

$$\text{var}(\hat{T}_\tau) = \underbrace{\sum_{j=1}^{J_\tau} \sum_{j'=1}^{J_\tau} \frac{\pi_{jj'}^{A_2} - \pi_j^{A_2} \pi_{j'}^{A_2}}{\pi_j^{A_2} \pi_{j'}^{A_2}} Z_j Z_{j'}}_{V_{A_2}^\tau} + \underbrace{\text{E}_{s_p^{A_2,t_\tau}} \left\{ \sum_{j \in s_p^{A_2,t_\tau}} \sum_{j' \in s_p^{A_2,t_\tau}} \frac{1}{\pi_j^{A_2}} \frac{1}{\pi_{j'}^{A_2}} \text{cov}_{jj'} \right\}}_{V_{\text{CNR}}^\tau}, \tag{5.8}$$

where J_τ designates the number of persons present in the population during the year in which $s_p^{A_2,t_\tau}$ is drawn. The first term $V_{A_2}^\tau$ is the portion of the variance due to the mechanism for selecting the longitudinals in $s_p^{A_2,t_\tau}$, while the second term V_{CNR}^τ is the portion due to households' non-response to the grid and then to the household questionnaire in year t, which constitutes cluster non-response.

To obtain an estimator of the variance of $\hat{T}_\tau$, we adapt the variance estimation formula (8.37) of Lavallée (2002, page 154). The following differences should be noted. First, we are ignoring the fact that in practice, the response probabilities will have to be estimated, whereas Lavallée (2002) takes this into account. Second, the estimation method proposed by Lavallée (2002) provides biased estimates, even when it is applied in a situation where the response probabilities are known. Consequently, we have adapted his method so as to obtain an unbiased estimator of the variance. To justify our approach, we first explain below the bias obtained by applying the method of Lavallée (2002) in a case where response probabilities are known. His method consists in estimating V_{CNR}^{τ} by an unbiased estimator $\hat{V}_{\mathrm{CNR}}^{\tau}$ then estimating $V_{A_2}^{\tau}$ by $\hat{V}_1^{\tau} = \tilde{V}_{A_2}^{\tau}(\hat{Z}_1, \hat{Z}_2, \ldots)$ where:

$$\tilde{V}_{A_2}^{\tau}(Z_1, Z_2, \ldots) = \sum_{j \in s_p^{A_2, t_\tau}} \sum_{j' \in s_p^{A_2, t_\tau}} \frac{\pi_{jj'}^{A_2} - \pi_j^{A_2}\pi_{j'}^{A_2}}{\pi_j^{A_2}\pi_{j'}^{A_2}} \frac{1}{\pi_{jj'}^{A_2}} Z_j Z_{j'} \tag{5.9}$$

is the Horvitz-Thompson estimator of the variance of

$$\tilde{T}_\tau = \sum_{j \in s_p^{A_2, t_\tau}} \frac{1}{\pi_j^{A_2}} Z_j. \tag{5.10}$$

This leads to the variance estimator:

$$\widehat{\mathrm{var}}_L(\hat{T}_\tau) = \hat{V}_1^{\tau} + \hat{V}_{\mathrm{CNR}}^{\tau}. \tag{5.11}$$

The use of $\hat{V}_1^{\tau}$, which is constructed by replacing the Z_j vales that appear in (5.9) by $\hat{Z}_j$, is motivated by the fact that the Z_j values are not known for all the longitudinals j in $s_p^{A_2, t_\tau}$, but only for the longitudinals who, in the survey year, belong to a household k that responded to the questionnaire, that is, a household k belonging to $s_m^{B,\tau}$. The use of $\hat{Z}_j$ values makes it possible to assign more weight to the longitudinals of $s_p^{A_2, t_\tau}$ for which the Z_j values are known. The problem is that the estimator $\hat{V}_1^{\tau}$ thus constructed does not provide an unbiased estimate of $V_{A_2}^{\tau}$. This may easily be seen by observing what happens for the diagonal terms: for any longitudinal j belonging to household k during the survey year, the quantity Z_j^2 appearing in (5.9) is replaced by $(\hat{Z}_j)^2 = Z_j^2 / (q_k^b q_k^c)^2 \mathbf{1}_{\{k \in s_m^{B,\tau}\}}$ while a weight increase of only a factor of $1 / (q_k^b q_k^c)$ is probably

a better choice. The same type of problem occurs for the product $Z_j Z_{j'}$ when longitudinal j and j' belong to the same household during the survey year. More specifically, we have, for all longitudinals j and j' belonging to $s_p^{A_2,t_\tau}$:

$$\mathrm{E}(\hat{Z}_j \hat{Z}_{j'} \mid s_p^{A_2,t_\tau}) = Z_j Z_{j'} + \mathrm{cov}_{jj'} \tag{5.12}$$

which implies

$$\mathrm{E}[\hat{V}_1^\tau] - V_{A_2}^\tau = \mathrm{E}_{s_p^{A_2,t_\tau}} \left\{ \mathrm{E}\left[\tilde{V}_{A_2}^\tau(\hat{Z}_1, \hat{Z}_2, \ldots) \mid s_p^{A_2,t_\tau} \right] \right\} - V_{A_2}^\tau \tag{5.13}$$

$$= \mathrm{E}_{s_p^{A_2,t_\tau}} \left[\sum_{j \in s_p^{A_2,t_\tau}} \sum_{j' \in s_p^{A_2,t_\tau}} \frac{\pi_{jj'}^{A_2} - \pi_j^{A_2}\pi_{j'}^{A_2}}{\pi_j^{A_2}\pi_{j'}^{A_2}} \frac{1}{\pi_{jj'}^{A_2}} \mathrm{cov}_{jj'} \right]. \tag{5.14}$$

Since on the other hand $\hat{V}_{\mathrm{CNR}}^\tau$ is an unbiased estimator of V_{CNR}^τ, we have:

$$\mathrm{B}^\tau = E\left[\widehat{\mathrm{var}}_L(\hat{T}_\tau) \right] - \mathrm{var}(\hat{T}_\tau) \tag{5.15}$$

$$= \mathrm{E}_{s_p^{A_2,t_\tau}} \left[\sum_{j \in s_p^{A_2,t_\tau}} \sum_{j' \in s_p^{A_2,t_\tau}} \frac{\pi_{jj'}^{A_2} - \pi_j^{A_2}\pi_{j'}^{A_2}}{\pi_j^{A_2}\pi_{j'}^{A_2}} \frac{1}{\pi_{jj'}^{A_2}} \mathrm{cov}_{jj'} \right]. \tag{5.16}$$

The bias B^τ depends in particular on the term $\mathrm{cov}_{jj'}$ defined by (5.7), and hence on the probabilities q_k^b and q_k^c of responding to the grid and the questionnaire in the survey year. Consider the simple case of the panel responding in the first wave, $\tau = 1$, in which the composition of the households has not yet begun to evolve. The quantity $\mathrm{cov}_{jj'}$ defined by (5.7) is positive if longitudinal j and j' belong to the same household k, and is otherwise nil. Also, for all longitudinals j and j' belonging to the same household k, we have, in accordance with the relation (2.4):

$$\frac{\pi_{jj'}^{A_2} - \pi_j^{A_2}\pi_{j'}^{A_2}}{\pi_{jj'}^{A_2}} = 1 - \pi_k^{A_2}. \tag{5.17}$$

Since the latter quantity is also positive, the expression of bias given by formula (5.16) means that B^1 is positive. On the other hand, the probabilities of inclusion $\pi_k^{A_2}$ are very low, and

therefore $1 - \pi_k^{A_2} \simeq 1$. Using (5.16) and (5.17), we are therefore able to make the following approximation:

$$\mathrm{B}^1 \simeq \mathrm{E}_{s_p^{A_2,t_1}} \left[\sum_{j \in s_p^{A_2,t_1}} \sum_{j' \in s_p^{A_2,t_1}} \frac{1}{\pi_j^{A_2}} \frac{1}{\pi_{j'}^{A_2}} \mathrm{cov}_{jj'} \right] = V_{\mathrm{CNR}}^1, \tag{5.18}$$

where, as noted above, V_{CNR}^{τ} is defined by (5.8) and corresponds in the first wave to the portion of the variance due to the non-response observed between the grid and the questionnaire. Consequently, the estimator $\widehat{\mathrm{var}}_L(\hat{T}_1)$ overestimates the variance of $\hat{T}_1$ and the error committed is of the order of magnitude of V_{CNR}^1. The bias may be relatively large if the probabilities of response to the questionnaire are low. As regards the other waves, the quantity $(\pi_{jj'}^{A_2} - \pi_j^{A_2}\pi_{j'}^{A_2}) / \pi_{jj'}^{A_2}$ that appears in (5.16) depends on the households k_1 and k_1' to which the longitudinals j and j' belonged during the year of their selection, and it is no longer easy to obtain an order of magnitude of the bias B^{τ}.

We introduce a term correcting the bias B^{τ} and we give our variance estimation formula in proposition 1, below. Keeping in mind that m_k designates all persons who comprise household k during survey year t (*cf.* page 11), let $\tilde{m}_k$ be the set, of cardinal L_k, consisting of the longitudinals j belonging to m_k.

Proposition 1: An unbiased estimate of the variance of $\hat{T}_{\tau}$ is given by

$$\widehat{\mathrm{var}}(\hat{T}_{\tau}) = \hat{V}_1^{\tau} + \hat{V}_2^{\tau}, \tag{5.19}$$

where

$$\hat{V}_1^{\tau} = \tilde{V}_{A_2}^{\tau}(\hat{Z}_1, \hat{Z}_2, \ldots) = \sum_{j \in s_p^{A_2,t_{\tau}}} \sum_{j' \in s_p^{A_2,t_{\tau}}} \frac{\pi_{jj'}^{A_2} - \pi_j^{A_2}\pi_{j'}^{A_2}}{\pi_j^{A_2}\pi_{j'}^{A_2}} \frac{1}{\pi_{jj'}^{A_2}} \hat{Z}_j \hat{Z}_{j'}$$

and

$$\hat{V}_2^{\tau} = \sum_{k \in s_m^{B,\tau}} \sum_{j,j' \in \tilde{m}_k} \frac{1}{\pi_{jj'}^{A_2}} \left(\frac{e_k'}{L_k + P_k} \right)^2 \frac{1 - q_k^b q_k^c}{(q_k^b q_k^c)^2}.$$

The demonstration of proposition 1 is provided in Appendix B.

Note 3: The estimator $\widehat{\text{var}}(\hat{T}_\tau) = \hat{V}_1^\tau + \hat{V}_2^\tau$ is the sum of two biased estimators whose biases are brought into balance by construction, with the result that $\widehat{\text{var}}(\hat{T}_\tau)$ gives an unbiased estimate of the variance of $\hat{T}_\tau$.

Note 4: Proposition 1 is based on the assumption that the values e_k and the response probabilities are known, which enables us to conclude that the estimator $\widehat{\text{var}}(\hat{T}_\tau)$ given by formula (5.19) is unbiased. In practice, these quantities must be estimated. The consequence of this is that the estimator of variance thus obtained is no longer unbiased but only asymptotically unbiased, provided that the non-response models can be considered correct and that their parameters are estimated by an appropriate method.

6 Final formula

The term $\hat{V}_2^\tau$ that appears in proposition 1 contains a double sum, but the latter does not pose a problem for operational purposes. The fact is that the proposition contains very few terms, since it applies only to the individuals in a single household. On the other hand, the expression $\hat{V}_1^\tau = \tilde{V}_{A_2}^\tau(\hat{Z}_1, \hat{Z}_2, \ldots)$ must be transformed to make it easier to calculate. We therefore begin by giving another expression of the term $\tilde{V}_{A_2}^\tau(Z_1, Z_2, \ldots)$ defined by formula (5.9). For this, we observe that

$$\tilde{T}_\tau = \sum_{j \in s_p^{A_2, t_\tau}} \frac{1}{\pi_j^{A_2}} Z_j = \sum_{k_1 \in s_m^{A_2, t_\tau}} \frac{1}{\pi_{k_1}^{A_2}} T_{k_1}, \tag{6.1}$$

where

$$T_{k_1} = \sum_{j \in m_{k_1}} Z_j. \tag{6.2}$$

Keep in mind that the selection of $s_m^{A_2, t_\tau}$ results from a design stratified by major region followed by a non-response stage modelled on a Poisson design on the households of wave 1 (*cf.* Section 2). Giving a simple expression of the Horvitz-Thompson estimator of variance of a total for this very

classical design is a problem that has already been widely studied, particularly in connection with the POULPE software for computing precision (*cf*. Caron, Deville and Sautory 1998, page 13). To give such an expression, we introduce the following notations. For each of the seven strata h, we let N_h denote the number of households that it contains and n_h the number of households selected. For any household $k_1 \in s_m^{A_1,t_\tau}$, we have:

$$T_{k_1}^* = \begin{cases} \dfrac{T_{k_1}}{q_{k_1}^a} & \text{if } k_1 \in s_m^{A_2,t_\tau} \\ 0 & \text{otherwise.} \end{cases} \tag{6.3}$$

Also, let us assume that for any h:

$$(s_h^*)^2 = \frac{1}{n_h - 1} \sum_{k_1 \in \{s_m^{A_1,t_\tau} \cap h\}} (T_{k_1}^* - \overline{T}_h^*)^2 = \frac{1}{n_h - 1} \left[\sum_{k_1 \in \{s_m^{A_2,t_\tau} \cap h\}} \left(\frac{T_{k_1}}{q_{k_1}^a} \right)^2 \right] - \frac{n_h}{n_h - 1} (\overline{T}_h^*)^2 \tag{6.4}$$

where

$$\overline{T}_h^* = \frac{1}{n_h} \sum_{k_1 \in \{s_m^{A_1,t_\tau} \cap h\}} T_{k_1}^* = \frac{1}{n_h} \sum_{k_1 \in \{s_m^{A_2,t_\tau} \cap h\}} \frac{T_{k_1}}{q_{k_1}^a}.$$

According to Caron *et al*. (1998, page 13), the term $\tilde{V}_{A_2}^{\tau}(Z_1, Z_2, \ldots)$ can be written here [see also formula (11.12) of Särndal and Lundström 2005):

$$\begin{aligned} \tilde{V}_{A_2}^{\tau}(Z_1, Z_2, \ldots) = & \sum_{k_1 \in s_m^{A_1,t_\tau}} \sum_{k_1' \in s_m^{A_1,t_\tau}} \frac{\pi_{k_1 k_1'}^{A_1} - \pi_{k_1}^{A_1} \pi_{k_1'}^{A_1}}{\pi_{k_1}^{A_1} \pi_{k_1'}^{A_1}} \frac{1}{\pi_{k_1 k_1'}^{A_1}} T_{k_1}^* T_{k_1'}^* - \sum_{k_1 \in s_m^{A_2,t_\tau}} \frac{1 - \pi_{k_1}^{A_1}}{(\pi_{k_1}^{A_1})^2} \frac{1 - q_{k_1}^a}{(q_{k_1}^a)^2} T_{k_1}^2 \\ & + \sum_{k_1 \in s_m^{A_2,t_\tau}} \frac{1 - q_{k_1}^a}{(q_{k_1}^a)^2} \left(\frac{T_{k_1}}{\pi_{k_1}^{A_1}} \right)^2. \end{aligned} \tag{6.5}$$

By grouping the last two terms and using that the sampling design of $s_m^{A_1,t_\tau}$ is a stratified design, we obtain the following simple expression for $\tilde{V}_{A_2}^{\tau}(Z_1, Z_2, \ldots)$:

$$\tilde{V}_{A_2}^{\tau}(Z_1, Z_2, \ldots) = \sum_{h=1}^{7} \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h} \right) (s_h^*)^2 + \sum_{k_1 \in s_m^{A_2,t_\tau}} \frac{1 - q_{k_1}^a}{(q_{k_1}^a)^2} \frac{1}{\pi_{k_1}^{A_1}} T_{k_1}^2. \tag{6.6}$$

Let $\hat{T}_{k_1}$ and $\hat{s}_h^*$ denote the estimators obtained by replacing the variables Z_j by $\hat{Z}_j$ in formulas (6.2) and (6.4). Relation (6.6), combined with formula (4.6) and proposition 1, makes it possible to obtain the final formula below for the estimate of the variance of complex estimator $\hat{\Theta}$:

$$\widehat{\text{var}}(\hat{\Theta}) \simeq \sum_{\tau=1}^{4} \left[\hat{V}_1^{\tau} + \hat{V}_2^{\tau} \right] = \underbrace{\sum_{\tau=1}^{4} \hat{V}_{1.1}^{\tau}}_{\hat{V}_{1.1}} + \underbrace{\sum_{\tau=1}^{4} \hat{V}_{1.2}^{\tau}}_{\hat{V}_{1.2}} + \underbrace{\sum_{\tau=1}^{4} \hat{V}_2^{\tau}}_{\hat{V}_2}, \tag{6.7}$$

where

$$\hat{V}_{1.1}^{\tau} = \sum_{h=1}^{7} \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h} \right) (\hat{s}_h^*)^2 \tag{6.8}$$

$$\hat{V}_{1.2}^{\tau} = \sum_{k_1 \in s_m^{A_2, t_\tau}} \frac{1 - q_{k_1}^a}{(q_{k_1}^a)^2} \frac{1}{\pi_{k_1}^{A_1}} (\hat{T}_{k_1})^2 \tag{6.9}$$

$$\hat{V}_2^{\tau} = \sum_{k \in s_m^{B,\tau}} \sum_{j, j' \in \tilde{m}_k} \frac{1}{\pi_{jj'}^{A_2}} \left(\frac{\hat{e}'_k}{L_k + P_k} \right)^2 \frac{1 - q_k^b q_k^c}{(q_k^b q_k^c)^2}. \tag{6.10}$$

Note 5: The variance estimation formula (6.7) always provides positive estimates. Also, the three terms that comprise it can be programmed very easily.

7 Numerical application and discussion

We will now present a numerical application to illustrate the results developed in the previous sections. Up to now, we have assumed that the response probabilities occurring in the different phases of the survey are known, but in practice, they must be estimated. As regards non-response to the grid in the first wave, the estimation of the response probabilities $q_{k_1}^a$ defined by (2.1) use information on non-respondent households collected via a survey of non-response. Sample $s_m^{A_1, t_\tau}$ of the households contacted during the first wave in year t_τ is divided into homogeneous response groups within which estimates $\hat{q}_{k_1}^a$ of response probabilities $q_{k_1}^a$ are calculated. The factors q_k^c appearing in (3.4), which represent the probabilities of response between the grid and the questionnaire in the survey year, are also estimated within homogeneous response groups. This time it is households that responded to the grid in the survey year that are allocated to homogeneous

response groups. Let $\hat{q}_k^c$ denote the estimate of response probability q_k^c thus obtained. It is slightly more problematic to estimate the factors q_j^b appearing in (3.2), which represent the probabilities of response between the grid of the first wave and that of the survey year. The following two options can be considered. The first option is to create homogeneous response groups within the set of households that survey personnel attempted to contact in the survey year. We are then faced with the problem that there is very little information by which to constitute these groups. The fact is that we do not know the current composition of the households that survey personnel did not manage to recontact or who refuse to fill out the grid again. The second option is to estimate the q_j^b values within the homogeneous response groups defined on the basis of the longitudinals of $s_p^{A_2,t_\tau}$. This is not entirely consistent with the response mechanism, but it allows us to use all the information collected regarding the longitudinals in previous waves. Since the first official processing of the survey, the choice has focused on the second option (see Graf 2008). Based on this estimation method, the quantities $\hat{q}_j^b$ are not all equal within a given household k for the survey year, which poses a problem when it comes to applying the variance estimation formulas developed in previous sections. This problem can circumvented based on the following finding. Let $\tilde{w}_k^p$ denote the shared weight obtained from (3.3) and from the estimation of response probabilities $q_{k_1}^a$ and q_j^b:

$$\tilde{w}_k^p = \frac{1}{L_k + P_k} \sum_{j=1}^{L_k} \left(\tilde{w}_j^{A_2} \frac{1}{\hat{q}_j^b} \right), \tag{7.1}$$

where $\tilde{w}_j^{A_2} = 1 / (\pi_{k_1}^{A_1} \hat{q}_{k_1}^a)$ for any longitudinal j belonging to household k_1 in the first wave. The shared weight $\tilde{w}_k^p$ of household k can be rewritten as follows:

$$\tilde{w}_k^p = \frac{1}{L_k + P_k} \sum_{j=1}^{L_k} \left(\tilde{w}_j^{A_2} \frac{1}{\hat{q}_j^b} \right) = \frac{1}{L_k + P_k} \sum_{j=1}^{L_k} \left(\tilde{w}_j^{A_2} \frac{1}{\tilde{q}_j^b} \right) \tag{7.2}$$

where, for all longitudinal j in household k, the quantities $\tilde{q}_j^b$ are equal to a common value $\tilde{q}_k^b$ defined by

$$\tilde{q}_k^b = \frac{\sum_{j=1}^{L_k} \tilde{w}_j^{A_2}}{\sum_{j=1}^{L_k} \left(\tilde{w}_j^{A_2} \frac{1}{\hat{q}_j^b} \right)}. \qquad (7.3)$$

Consequently, whether we estimate the quantities q_j^b by $\hat{q}_j^b$ or by $\tilde{q}_j^b$, we obtain exactly the same shared weight $\tilde{w}_k^p$, and therefore the same final weight and the same estimates of poverty indexes. On the basis of this finding, we obtain an approximation of the precision of $\hat{\Theta}$ by replacing, in the variance estimation formulas, the unknown quantities $q_{k_1}^a, q_k^b, q_k^c$ by $\hat{q}_{k_1}^a, \tilde{q}_k^b, \hat{q}_k^c$. The e_k values are also replaced by the $\hat{e}_k$ values given by (4.3). So as not to unnecessarily complicate the notation, we use the same notation for the variance estimators calculated on the basis of the unknown parameters and those calculated on the basis of estimates of them. If the estimators of response probabilities are convergent, the estimator $\widehat{\text{var}}(\hat{\Theta})$ that we propose in formula (6.7) is asymptotically unbiased. The asymptotic nature of this property is due to the effect of the estimation of response probabilities as well as to the approximations made in the course of linearizing, taking calibration into account and obtaining formula (4.6).

The calculations presented in this section were made on the basis of the 2009 SILC-Switzerland survey. SILC did not begin in Switzerland until 2007. Thus, by 2009, the survey was not yet complete, as it was composed of only three panels rather than four. However, this has no effect on the methodology described in the previous sections. In 2009, 7,372 households, or 17,561 individuals, agreed to participate in the survey. Table 7.1 shows the response rates obtained in the different stages of the survey for each of the panels.

Table 7.1
Response rate for different stages of the survey

Response rate	Panel 07	Panel 08	Panel 09
In the grid in wave 1	0.688	0.694	0.687
Between the grid in wave 1 and the current wave	0.803	0.834	1.000
Between the grid and the household questionnaire for the current wave	0.966	0.951	0.942

Based on the 2009 survey, the poverty rates of the different population subgroups were calculated, analyzed and commented on by the Swiss Federal Statistical Office (FSO 2010). Table 7.2 shows the different measures of precision for a few of the most commonly used inequality indexes: at-risk-of-poverty index, quintile share ratio, Gini coefficient and relative median at-risk-of-poverty gap (RMPG). For each indicator $\hat{\Theta}$ considered, Table 7.2 contains the following measures:

$$\text{SD} = \sqrt{\widehat{\text{var}}(\hat{\Theta})}, \quad \text{CV} = \frac{\text{SD}}{\hat{\Theta}}, \quad \text{SD}_{1.1} = \sqrt{\hat{V}_{1.1}}, \quad \Delta_1 = \frac{\text{SD}_{1.1}}{\text{SD}} - 1,$$

where the variances $\widehat{\text{var}}(\hat{\Theta})$ and $\hat{V}_{1.1}$ are given by formula (6.7). Let $\widehat{\text{var}}_L(\hat{\Theta})$ denote the variance estimator obtained by the method of Lavallée (2002):

$$\widehat{\text{var}}_L(\hat{\Theta}) = \sum_{\tau=1}^{3} \widehat{\text{var}}_L(\hat{T}_\tau), \tag{7.4}$$

where $\widehat{\text{var}}_L(\hat{T}_\tau)$ is defined by equation (5.11). The measures,

$$\text{SD}_L = \sqrt{\widehat{\text{var}}_L(\hat{\Theta})}, \text{CV}_L = \frac{\text{SD}_L}{\hat{\Theta}}, \Delta_2 = \frac{\text{SD}_L}{\text{SD}} - 1,$$

are also shown in Table 7.2.

Table 7.2
Precisions estimated on the basis of 2009 data for different indicators

Indicator	$\hat{\Theta}$	SD	CV	$\text{SD}_{1.1}$	Δ_1	SD_L	CV_L	Δ_2
At-risk-of-poverty rate								
Total population	14.6%	0.581	4.0%	0.580	-0.61%	0.632	4.3%	8.9%
0-17 years	18.3%	1.310	7.2%	1.310	-0.60%	1.426	7.9%	8.8%
18-24 years	11.9%	1.302	10.9%	1.301	-0.61%	1.388	11.6%	6.6%
25-49 years	10.7%	0.585	5.5%	0.585	-0.62%	0.645	6.1%	10.2%
50-64 years	9.9%	0.752	7.6%	0.752	-0.63%	0.828	8.4%	10.1%
65 years and over	26.4%	1.291	4.8%	1.291	-0.62%	1.442	5.4%	11.6%
Quintile share ratio	4.4	0.109	2.4%	0.109	-0.55%	0.118	2.7%	8.9%
Gini coefficient	***[a]	***	1.7%	***	-0.56%	***	1.8%	8.0%
RMPG	***[a]	***	5.8%	***	-0.60%	***	6.4%	9.7%

[a] We do not report the value of these indicators, since they have not yet been published by the Swiss Federal Statistical Office.

Table 7.2 lends itself to comparisons between the different measures of precision, leading to the following observations. First, for the indicators considered, almost the entire estimate of the standard deviation SD of indicator $\hat{\Theta}$ is contained in the term $\mathrm{SD}_{1.1}$. The relative difference Δ_1 between SD and $\mathrm{SD}_{1.1}$ varies between -0.55% for the quintile share ratio and -0.63% for the at-risk-of-poverty rate for persons aged 50 to 64. Thus, by calculating only the dominant term $\mathrm{SD}_{1.1}$, we generally obtain an excellent approximation of the estimate of the standard deviation of $\hat{\Theta}$. This term is extremely simple to program. On the other hand, the method of estimating the variance of $\hat{\Theta}$ proposed by Lavallée (2002) – which, as formula (5.14) shows, is not unbiased even when the response probabilities are known – leads here to slightly higher estimates of precision than with the estimation method that we propose. The relative difference Δ_2 between SD and SD_L ranges between 6.6% for the at-risk-of-poverty rate for persons aged 18-24 and 11.6% for the rate for those aged 65 and over. From formulas (5.7) and (5.16), it emerges that the estimator SD_L is asymptotically unbiased and equal to SD when the probabilities of responding to the grid and the household questionnaire q_k^b and q_k^c in the survey year are equal to 1. While the rates of response to the grid and the household questionnaire in the survey year are fairly high (*cf.* Table 7.1), the differences between SD and SD_L are non-negligible.

Conclusion

We have proposed a variance estimator for poverty and social exclusion indicators, one that takes account of the non-linearity of the estimators, total non-response in different stages of the survey, indirect sampling and calibration. Ideally, the effects of imputations should also be taken into account. However, it should be noted that our approach is compatible with the requirements of Eurostat (2010), to whom many European countries provide only an approximation of the variance due to sampling and total non-response, along with minimal indications on imputations such as the percentage of imputed values. We have modified the method proposed by Lavallée (2002) for

estimating variance in the presence of non-response after weight sharing in order to correct the bias that this method induces. Our estimator is always positive in the case of the SILC-Switzerland survey, and it consists of three terms that are quite simple to program. Also, in calculating only the first of these three terms, we obtain an excellent approximation of variance.

Acknowledgements

I wish to thank Yves Tillé and Lionel Qualité of the Université de Neuchâtel for the productive discussions, advice and attentive review. Warms thanks are also extended to Eric Graf, Johan Pea, Thomas Christin and Stéphane Fleury of the Swiss Federal Statistical Office, who provided the information needed for the project to go smoothly. I would like to thank Philippe Eichenberger and Monique Graf of the Federal Statistical Office for their support. Finally, I am very grateful to the judges and the associate editor for having taken the time to carefully review this work. Their comments led to many improvements.

Appendix A

Linearization formulas

What follows is a review of the linearization formulas of Osier (2009) in the case of the four poverty indicators examined in Section 7. Let $\mathcal{U}$ be the population of individuals present in the survey year. In accordance with the notations used in Section 5, J_1 denotes the size of $\mathcal{U}$, since $\mathcal{U}$ corresponds to the target population of the panel responding in wave 1. However, to simplify the notations in this appendix, the size of $\mathcal{U}$ is simply denoted by J. Let R_j denote the equivalent income of individual $j \in \mathcal{U}$ used to calculate the poverty indicators, and the following notation is used for any $x \in \mathbb{R}$:

$$F(x) = \frac{1}{J} \sum_{j \in \mathcal{U}} \mathbf{1}_{\{R_j \leq x\}}, \tag{A.1}$$

where $\mathbf{1}_{\{R_j \leq x\}}$ is an indicator variable that equals 1 if the income R_j of individual j is less than or equal to x, and 0 otherwise. Linearization formulas for the poverty indicators are first obtained by assuming that F is derivable and that F' is non-nil. Since this is not the case, we get around this problem by approaching F by the function F_K defined, for any $x \in \mathbb{R}$, as:

$$F_K(x) = \int F(z)K(x, z)dz, \tag{A.2}$$

where

$$K(x, z) = \frac{1}{h\sqrt{2\pi}} \exp\left[-\frac{(x - z)^2}{2h^2} \right], \tag{A.3}$$

with the parameter h being a smoothing parameter.

1. At-risk-of-poverty rate

   The at-risk-of-poverty threshold, ARPT, is calculated on the basis of the median income MED of the population $\mathcal{U}$:

   $$\text{ARPT} = 0.6 \times \text{MED}. \tag{A.4}$$

   The at-risk-of-poverty rate, ARPR, is defined as follows:

   $$\text{ARPR} = \frac{1}{J} \sum_{j \in \mathcal{U}} \mathbf{1}_{\{R_j \leq \text{ARPT}\}}. \tag{A.5}$$

   The linearized ℓ_j that appears in the approximation of the variance of the at-risk-of-poverty rate estimator, given in formula (4.1), is written as follows for each individual j:

   $$\ell_j = \frac{1}{J}\left[\mathbf{1}_{\{R_j \leq \text{ARPT}\}} - \text{ARPR} \right] - \frac{0.6}{J} \cdot \frac{F_K'(\text{ARPT})}{F_K'(\text{MED})} \cdot [\mathbf{1}_{\{R_j \leq \text{MED}\}} - 0.5]. \tag{A.6}$$

   In Section 7, the poverty rate was also estimated within different sub-populations. Formula (A.6) is easily generalized in the case of sub-populations and is not reviewed here.

2. Quintile share ratio

   Let γ_+ be the quantile of order 0.8 and let γ_- be the quantile of order 0.2, and then let:

   $$R = \sum_{j \in \mathcal{U}} R_j \text{ and } S(x) = \sum_{j \in \mathcal{U}} R_j \mathbf{1}_{\{R_j \leq x\}}. \tag{A.7}$$

The quintile share ratio, denoted QSR, is defined as follows:

$$\text{QSR} = \frac{R - S(\gamma_+)}{S(\gamma_-)}. \tag{A.8}$$

Similar to what was done for the function F, we approach the function S by the derivable function S_K defined as:

$$S_K(x) = \int S(z)K(x, z)dz. \tag{A.9}$$

The linearized ℓ_j that appears in the approximation of the variance of the quintile share ratio estimator, which was given in formula (4.1), is written as follows for each individual j:

$$\ell_j = \frac{S(\gamma_-)\left[R_j - Q_j(\gamma_+)\right] - \left[R - S(\gamma_+)\right]Q_j(\gamma_-)}{\left[S(\gamma_-)\right]^2}, \tag{A.10}$$

where the quantity $Q_j(\gamma)$ is defined, for any quantile γ of order α, as follows:

$$Q_j(\gamma) = R_j \mathbf{1}_{\{R_j \le \gamma\}} - \frac{S'_K(\gamma)}{JF_K(\gamma)}[\mathbf{1}_{\{R_j \le \gamma\}} - \alpha]. \tag{A.11}$$

3. Gini coefficient

The Gini coefficient, denoted G, is defined as:

$$\text{G} = \frac{2M - R}{JR} - 1, \tag{A.12}$$

where

$$M = \sum_{j \in \mathcal{U}} R_j \left(\sum_{j' \in \mathcal{U}} \mathbf{1}_{\{R_{j'} \le R_j\}} \right) \tag{A.13}$$

and R is defined by (A.7). The linearized ℓ_j appearing in the approximation of the variance of the Gini coefficient estimator, given in formula (4.1), is written as follows for each individual j:

$$\ell_j = \frac{JR(2U_j - R_j) - (2M - R)(R + JR_j)}{(JR)^2}, \tag{A.14}$$

where the quantity U_j is defined by

$$U_j = \sum_{j' \in \mathcal{U}} R_{j'} \mathbf{1}_{\{R_j \le R_{j'}\}} + R_j \sum_{j' \in \mathcal{U}} \mathbf{1}_{\{R_{j'} \le R_j\}}. \tag{A.15}$$

4. Relative median at-risk-of-poverty gap

   Relative median at-risk-of-poverty gap, denoted by RMPG, is defined by:

$$\mathrm{RMPG} = \frac{\mathrm{ARPT} - \mathrm{MED}^p}{\mathrm{ARPT}}, \tag{A.16}$$

   where as previously noted, ARPT designates the poverty threshold and where MED^p is the median income calculated for individuals with an income below the poverty threshold. The linearized ℓ_j appearing in the approximation of the variance of the RMPG estimator, given in formula (4.1), is written as follows for each individual j:

$$\ell_j = -\frac{\mathrm{ARPT} \cdot Y_j - \mathrm{MED}^p \cdot W_j}{(\mathrm{ARPT})^2}, \tag{A.17}$$

   where W_j is defined by

$$W_j = -\frac{0.6}{F_K'(\mathrm{MED})} \frac{1}{J} [\mathbf{1}_{\{R_j \leq \mathrm{MED}\}} - 0.5] \tag{A.18}$$

   and Y_j verifies the equation

$$\begin{aligned} F_K'(\mathrm{MED}^p) Y_j = {} & \frac{1}{2} \left\{ \frac{1}{J} \left[\mathbf{1}_{\{R_j \leq \mathrm{ARPT}\}} - F(\mathrm{ARPT}) \right] + F_K'(\mathrm{ARPT}) \cdot W_j \right\} \\ & - \frac{1}{J} \left[\mathbf{1}_{\{R_j \leq \mathrm{MED}^p\}} - F(\mathrm{MED}^p) \right]. \end{aligned} \tag{A.19}$$

Appendix B

Demonstration of proposition 1

The proof has four parts.

1. Definition of $\widehat{\mathrm{cov}}_{jj'}$

   For all longitudinals j and j' within $s_p^{A_2,t_\tau}$ and belonging to households k and k' respectively during the survey year, let

$$\widehat{\mathrm{cov}}_{jj'} = \begin{cases} 0 & \text{if } k \neq k' \\ \dfrac{1}{(L_k + P_k)^2} (e_k')^2 \dfrac{1 - q_k^b q_k^c}{(q_k^b q_k^c)^2} \mathbf{1}_{\{k \in s_m^{B,\tau}\}} & \text{if } k = k', \end{cases} \tag{B.1}$$

   with the result that

$$\mathrm{E}\left[\widehat{\mathrm{cov}}_{jj'} \mid s_p^{A_2,t_\tau} \right] = \mathrm{cov}_{jj'}. \qquad \text{(B.2)}$$

2. Estimation of $V_{A_2}^{\tau}$

   If the Z_j values were known for $s_p^{A_2,t_\tau}$, we could estimate $V_{A_2}^{\tau}$ without bias by $\tilde{V}_{A_2}^{\tau}(Z_1, Z_2, \ldots)$. Since this is not the case, we estimate $V_{A_2}^{\tau}$ without bias by:

$$\hat{V}_{A_2}^{\tau} = \sum_{j \in s_p^{A_2,t_\tau}} \sum_{j' \in s_p^{A_2,t_\tau}} \frac{\pi_{jj'}^{A_2} - \pi_j^{A_2}\pi_{j'}^{A_2}}{\pi_j^{A_2}\pi_{j'}^{A_2}} \frac{1}{\pi_{jj'}^{A_2}} \left[\hat{Z}_j \hat{Z}_{j'} - \widehat{\mathrm{cov}}_{jj'} \right].$$

   Indeed, for all longitudinals j and j' belonging to $s_p^{A_2,t_\tau}$, we have:

$$\mathrm{E}\left[\hat{Z}_j \hat{Z}_{j'} - \widehat{\mathrm{cov}}_{jj'} \mid s_p^{A_2,t_\tau} \right] = Z_j Z_{j'},$$

   which implies that

$$\mathrm{E}(\hat{V}_{A_2}^{\tau}) = \mathrm{E}_{s_p^{A_2,t_\tau}}\left[\mathrm{E}\left(\hat{V}_{A_2}^{\tau} \mid s_p^{A_2,t_\tau} \right) \right] = \mathrm{E}_{s_p^{A_2,t_\tau}}\left[\tilde{V}_{A_2}^{\tau}(Z_1, Z_2, \ldots) \right] = V_{A_2}^{\tau}.$$

3. Estimation of V_{CNR}^{τ}

   We can estimate V_{CNR}^{τ} by:

$$\hat{V}_{\mathrm{CNR}}^{\tau} = \sum_{j \in s_p^{A_2,t_\tau}} \sum_{j' \in s_p^{A_2,t_\tau}} \frac{1}{\pi_j^{A_2}} \frac{1}{\pi_{j'}^{A_2}} \widehat{\mathrm{cov}}_{jj'}.$$

4. Final formula

   We estimate $\mathrm{var}(\hat{T}_\tau)$ by:

$$\widehat{\mathrm{var}}(\hat{T}_\tau) = \hat{V}_{A_2}^{\tau} + \hat{V}_{\mathrm{CNR}}^{\tau}.$$

   After a few simplifications, we have:

$$\widehat{\mathrm{var}}(\hat{T}_\tau) = \tilde{V}_{A_2}^{\tau}(\hat{Z}_1, \hat{Z}_2, \ldots) + \sum_{j \in s_p^{A_2,t_\tau}} \sum_{j' \in s_p^{A_2,t_\tau}} \frac{1}{\pi_{jj'}^{A_2}} \widehat{\mathrm{cov}}_{jj'}. \qquad \text{(B.3)}$$

   Using (B.1), we can simplify the second term of the right member, so as to obtain formula (5.19) given in proposition 1.

References

Ardilly, P., and Lavallée, P. (2007). Weighting in rotating samples: The SILC survey in France. *Survey Methodology*, 33, 2, 131-137.

Caron, N., Deville, J. and Sautory, O. (1998). Estimation de données issues d'enquêtes : document méthodologique sur le logiciel POULPE. Technical report 9806, INSEE, Paris.

Clemenceau, A., and Museux, J. (2006). EU-SILC (Community Statistics on Income and Living Conditions): General presentation of the instrument. In *Proceedings of the EU-SILC Conference* (Helsinki, November 6 to 8, 2006).

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology*, 25, 2, 193-203.

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

Eurostat (2003). *Règlement (CE) No 1177/2003 du parlement européen et du conseil du 16 juin 2003 relatif aux statistiques communautaires sur le revenu et les conditions de vie (EU-SILC)*. Luxembourg, Official Journal of the European Union.

Eurostat (2010). 2008 comparative EU intermediate quality report. Technical report, European Commission, Luxembourg. http://circa.europa.eu/Public/irc/dsis/eusilc/library?l=/quality_assessment/comparative_quality_1/intermediate_version/_EN_1.0_&a=d.

FSO (2010). *Les conditions de vie en Suisse en 2009, Résultats de l'enquête sur les revenus et les conditions de vie* (*SILC*). Order No. 1192-0900. http://www.bfs.admin.ch/bfs/portal/fr/index/news/publikationen.Document.138922.pdf.

Graf, E. (2008). Pondérations du SILC pilote, SILC i vague 2, SILC ii vague 1, SILC i et SILC ii combinés. Technical report 338-0051, FSO, Neuchâtel. http://www.bfs.admin.ch/bfs/portal/de/index/infothek/erhebungen__quellen/methodenberichte.Document.105401.pdf.

Langel, M., and Tillé, Y. (2011). Statistical inference for the quintile share ratio. *Journal of Statistical Planning and Inference*, 141, 2976-2985.

Lavallée, P. (2002). *Le sondage indirect ou la méthode généralisée du partage des poids*. Paris: Ellipses.

Merkouris, T. (2001). Cross-sectional estimation in multiple-Panel household surveys. *Survey Methodology*, 27, 2, 171-181.

Osier, G. (2009). Variance estimation for complex indicators of poverty and inequality using linearization techniques. *Survey Research Methods*, 3, 167-195.

Särndal, C.-E., and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. New York: John Wiley & Sons, Inc.

Survey Methodology, June 2013
Vol. 39, No. 1, pp. 149-182
Statistics Canada, Catalogue No. 12-001-X

Combining cohorts in longitudinal surveys

Iván A. Carrillo and Alan F. Karr[1]

Abstract

A question that commonly arises in longitudinal surveys is the issue of how to combine differing cohorts of the survey. In this paper we present a novel method for combining different cohorts, and using all available data, in a longitudinal survey to estimate parameters of a semiparametric model, which relates the response variable to a set of covariates. The procedure builds upon the Weighted Generalized Estimation Equation method for handling missing waves in longitudinal studies. Our method is set up under a joint-randomization framework for estimation of model parameters, which takes into account the superpopulation model as well as the survey design randomization. We also propose a design-based, and a joint-randomization, variance estimation method. To illustrate the methodology we apply it to the Survey of Doctorate Recipients, conducted by the U.S. National Science Foundation.

Key Words: Superpopulation parameters; Joint-randomization inference; Replication variance estimation; Rotating panel surveys; Multi-cohort longitudinal surveys; Weighted Generalized Estimating Equations.

1 Introduction

The Survey of Doctorate Recipients (SDR) is a National Science Foundation (NSF) longitudinal survey whose design incorporates features of both repeated panels and rotating panels. The purpose of the survey is to study U.S. doctorate recipients in science, engineering, and health fields. It is conducted approximately every two years. A detailed description of the SDR can be found at NSF (2012). In this paper we restrict our attention to the data collected from 1995 through 2008 (7 waves).

At any particular wave a new cohort is selected. The new cohort consists of a sample of recent graduates (from the previous two years) selected from the Doctorate Records File, which is a database constructed mainly from the Survey of Earned Doctorates (http://www.nsf.gov/statistics/srvydoctorates/). The selected individuals are kept in the sample, *i.e.*, interviewed every two years, until the age of 75, while living in the U.S. during the survey reference week, and

1. Iván A. Carrillo and Alan F. Karr, National Institute of Statistical Sciences, 19 T.W. Alexander Drive, Research Triangle Park, NC 27709, U.S.A. E-mail: ivan@niss.org and karr@niss.org.

while not institutionalized. However, *not* all the sampled graduates satisfying these characteristics are retained forever. Some individuals, rather than entire cohorts, are dropped from the sample in order to a) include the new graduates in the new cohorts and b) maintain a relatively constant sample size across waves. In Section 2.2 we describe how the selection of the individuals who are dropped is made.

Survey weights for cross-sectional analyses of the SDR are already available, but not for longitudinal analyses. Rather than requiring a *new* longitudinal weight for *all* the data, the method proposed in this paper is able to use the existing cross-sectional weights for longitudinal analyses without ignoring any data. We concentrate on estimation of parameters of statistical models of the effect of covariates on a response of interest, but the method can also be used for estimation of finite population quantities (Carrillo and Karr 2012). We focus on analysis of the SDR, but our method is applicable to any fixed-panel, fixed-panel-plus-'births', repeated-panel, rotating-panel, split-panel, or refreshment sample survey, as long as for each wave there is a cross-sectional weight to represent the population of interest at that wave. See Smith, Lynn and Elliot (2009), Hirano, Imbens, Ridder and Rubin (2001), and Nevo (2003) for definitions of all these types of longitudinal sample designs.

The SDR is a hybrid of repeated-panel and rotating panel designs. It is not purely a repeated-panel design because of the removal of some subjects at each wave. It is not purely a rotating-panel design because entire panels (or cohorts) are *not* removed, only individuals; additionally, the composition of the finite population of interest changes over time, unlike in a rotating panel survey.

Diggle, Heagerty, Liang and Zeger (2002) and Hedeker and Gibbons (2006) point out that, with longitudinal studies, contrary to a cross-sectional study, it is possible to separate age effect (actual change within subjects over time) and cohort effect (difference between units at the beginning of the study period).

Hedeker and Gibbons (2006) also suggest that since longitudinal studies allow for the measurement of time-varying explanatory variables (covariates), the statistical inferences about dynamic relationship between the outcome on interest (response) and these covariates are much stronger than those based on cross-sectional studies.

When we are interested in the marginal mean of a variable, possibly conditionally on some covariates, and not in measuring change, a longitudinal study is not necessary; a cross-sectional study suffices. However, even in this case, a longitudinal study tends to be more powerful, because each subject serves as his or her own control for any unmeasured characteristics (Diggle, *et al.* 2002).

Our approach differs from the existing alternatives in the literature, which have some limitations for analysis of such data, and in particular for application to the SDR. For example, Berger (2004a) and Berger (2004b) go into detail about the estimation of change using rotating samples, but they assume that the composition of the finite population does not change over time, which is not the case of the SDR. This assumption does not hold in many other large-scale surveys. Also, the methodology proposed by Berger is not easily generalizable to more than two waves. Similarly, Qualité and Tillé (2008) also assume the finite population is fixed over time. Hirano, *et al.* (2001) and Nevo (2003) present different methods of estimation assuming a fixed-panel plus refreshment for attrition design, but also assume the finite population composition is fixed over time.

A time series approach is utilized by McLaren and Steel (2000) and Steel and McLaren (2007) to estimate change and trend with survey data. Although their approach allows for the incorporation of within-subject association in the point estimates, they do not consider covariates in their models (beyond the implicit time covariates). Also, they only discuss the estimation of change for continuous variables.

Another alternative for analyzing longitudinal data is to fix the finite population of interest, except perhaps for deaths, which could be allowed. Studies of this kind are those where there are

data available only for a single cohort. For example, Vieira and Skinner (2008), Carrillo, Chen and Wu (2010), and Carrillo, Chen and Wu (2011) show some alternatives for modeling with single-cohort survey data. However, to use these kinds of analyses with multi-cohort surveys, one needs to ignore some (or many) available data, for example those data from subjects who are not common to all waves. An example of a weighting procedure of this type can be found in Ardilly and Lavallée (2007).

Finally, the approach of Larsen, Qing, Zhou and Foulkes (2011) is appealing, in principle, because it is the way survey practitioners generally proceed. An initial weight is adjusted, among other things for calibration to known totals, in this case totals by survey wave. Nonetheless, for rotating panels this method is still in its infancy; there are some things that are not completely clear how to carry out. For example, it is not clear what the initial weight should be: a constant weight?, the earliest available weight?, the average of the available weights for each case?, or the latest available weight? Also, in the case of dropouts, as there exist in the SDR, the authors do not clarify how to carry out a nonresponse adjustment with this method. Even more, it is not clear why a nonresponse adjustment for dropouts at, say, wave 4 should have any influence on the observations at wave 3, as this methodology permits since there is a single weight for each subject. Additionally, the authors mention that they estimated standard errors, but they do not indicate how to take into account all the features of the sampling design, such as changes over time in the stratification and weighting adjustment classes of the SDR. Our method, on the other hand, utilizes only cross-sectional weights and variance estimation methods, which have been studied thoroughly in the literature and are readily available for the SDR.

The rest of the paper is organized as follows. In the next section we give a description of the SDR design. After that, in Section 3, we propose a novel approach for longitudinal analysis of marginal mean models with multi-cohort surveys. Then we present the application of the methodology to the SDR. Finally we offer a few discussion points in Section 5.

2 The SDR design

2.1 Finite population

The SDR finite population of interest can be represented as in Table 2.1. At wave 1, *i.e.*, the first time of interest, there is a finite set, $U_{1(1)} = U_1$, of $N_{1(1)} = N_1$ Ph.D. holders, either recent or not, who satisfy the requirements of the SDR.

Table 2.1
SDR finite population

j:	1		2		3		…		$J-1$		J
	$U_{1(1)}$	$\supseteq$	$U_{2(1)}$	$\supseteq$	$U_{3(1)}$	$\supseteq$	…	$\supseteq$	$U_{J-1(1)}$	$\supseteq$	$U_{J(1)}$
	$N_{1(1)}$	$\geq$	$N_{2(1)}$	$\geq$	$N_{3(1)}$	$\geq$	…	$\geq$	$N_{J-1(1)}$	$\geq$	$N_{J(1)}$
			$U_{2(2)}$	$\supseteq$	$U_{3(2)}$	$\supseteq$	…	$\supseteq$	$U_{J-1(2)}$	$\supseteq$	$U_{J(2)}$
			$N_{2(2)}$	$\geq$	$N_{3(2)}$	$\geq$	…	$\geq$	$N_{J-1(2)}$	$\geq$	$N_{J(2)}$
						$\ddots$			$\vdots$		$\vdots$
									$U_{J-1(J-1)}$	$\supseteq$	$U_{J(J-1)}$
									$N_{J-1(J-1)}$	$\geq$	$N_{J(J-1)}$
											$U_{J(J)}$
											$N_{J(J)}$
	U_1		U_2		U_3		…		U_{J-1}		U_J
	N_1		N_2		N_3		…		N_{J-1}		N_J

At wave 2 only a subset of the subjects in $U_{1(1)}$ still satisfy the SDR requirements; we call this subset, of $N_{2(1)}$ subjects, $U_{2(1)}$. In addition, there is a set of new, recent Ph.D. recipients, who have obtained their degree since wave 1, and also satisfy the other requirements of the survey. This set of new graduates in scope is called $U_{2(2)}$ and is of size $N_{2(2)}$. Therefore, at wave 2, there is a total of $N_2 = N_{2(1)} + N_{2(2)}$ subjects in the population of interest $U_2 = U_{2(1)} \cup U_{2(2)}$.

At the next wave, wave 3, the same process occurs. Some people in $U_{2(1)}$ leave the population of interest and there are only $N_{3(1)}$ left in $U_{3(1)}$. The same thing happens with the set $U_{2(2)}$; only a subset $U_{3(2)}$ of $N_{3(2)}$ among them still satisfy the requirements of the SDR. Additionally, there are $N_{3(3)}$ recent graduates entering the population of interest; this set is called $U_{3(3)}$. In total, the finite population of interest at wave 3 is $U_3 = U_{3(1)} \cup U_{3(2)} \cup U_{3(3)}$, with $N_3 = N_{3(1)} + N_{3(2)} + N_{3(3)}$ subjects.

This procedure, of thinning of old cohorts and adding new cohorts, continues until the last wave of interest, wave J. We notice that the finite population of interest changes at every wave due to two main reasons. Firstly, some of the subjects in the old cohorts are no longer in scope at the current wave, and they are not part of the current target population. Secondly, the recent graduates are added to the target population in the current wave. We denote by $j = 1, 2, \ldots, J$ the wave of interest (outside the parenthesis) and by $j' = 1, 2, \ldots, J$ the cohort to which a subject belongs (inside the parenthesis), and therefore $U_{j(j')} = U_{\text{wave(cohort)}}$.

2.2 Sampling

The sampling design of the SDR has a similar structure to the finite population and is depicted in Table 2.2. At wave 1, a (complex) sample $s_{1(1)} = s_1$ of $n_{1(1)} = n_1$ subjects is selected from within the N_1 elements in U_1. Each element i in s_1 is interviewed and its data collected; also, there is a design weight $w_{i1} = 1 / \pi_{i1}$ associated with it, which is the inverse of its inclusion probability at wave 1.

Table 2.2
SDR Sample

j :	**1**		**2**		**3**		**…**		**$J-1$**		**J**
	$s_{1(1)}$	$\supseteq$	$s_{2(1)}$	$\supseteq$	$s_{3(1)}$	$\supseteq$	$\cdots$	$\supseteq$	$s_{J-1(1)}$	$\supseteq$	$s_{J(1)}$
	$n_{1(1)}$	$\geq$	$n_{2(1)}$	$\geq$	$n_{3(1)}$	$\geq$	$\cdots$	$\geq$	$n_{J-1(1)}$	$\geq$	$n_{J(1)}$
			$s_{2(2)}$	$\supseteq$	$s_{3(2)}$	$\supseteq$	$\cdots$	$\supseteq$	$s_{J-1(2)}$	$\supseteq$	$s_{J(2)}$
			$n_{2(2)}$	$\geq$	$n_{3(2)}$	$\geq$	$\cdots$	$\geq$	$n_{J-1(2)}$	$\geq$	$n_{J(2)}$
					$s_{3(3)}$	$\supseteq$	$\cdots$	$\supseteq$	$s_{J-1(3)}$	$\supseteq$	$s_{J(3)}$
					$n_{3(3)}$	$\geq$	$\cdots$	$\geq$	$n_{J-1(3)}$	$\geq$	$n_{J(3)}$
							$\ddots$		$\vdots$		$\vdots$
									$s_{J-1(J-1)}$	$\supseteq$	$s_{J(J-1)}$
									$n_{J-1(J-1)}$	$\geq$	$n_{J(J-1)}$
											$s_{J(J)}$
											$n_{J(J)}$
	s_1		s_2		s_3		$\cdots$		s_{J-1}		s_J
	n_1		n_2		n_3		$\cdots$		n_{J-1}		n_J

At the second wave, the elements in $s_{1(1)}$ who are not in scope anymore are simply dropped from the frame (though their observations at wave 1 are kept), and a subsample $s_{2(1)}$, of size $n_{2(1)}$, of those still in scope is selected. Not all the members in $s_{1(1)}$ who are still in scope at wave 2 are retained in the sample; this is in order to be able to make up room for the sample of the new Ph.D. recipients and still maintain more or less the same sample size as in wave 1. A sample $s_{2(2)}$ of size $n_{2(2)}$ is selected from $U_{2(2)}$; people in $s_{2(2)}$ form the second cohort. The total sample at wave 2 is $s_2 = s_{2(1)} \cup s_{2(2)}$, which is of size $n_2 = n_{2(1)} + n_{2(2)}$, which is approximately equal to n_1. All the people in s_2 are interviewed at wave 2. The design weights at wave 2, $w_{i2} = 1 / \pi_{i2}$, are such that the sample s_2 represents the population of interest at wave 2, namely U_2.

The same procedure is repeated at each wave, till the last one (J), where a subsample of the remaining subjects from each of the previous $J - 1$ cohorts is selected, and a new sample (the new cohort) $s_{J(J)}$ of recent graduates is selected from $U_{J(J)}$. At the last wave, all people in $s_J = \bigcup_{j'=1}^{J} s_{J(j')}$ are interviewed and a design weight $w_{iJ} = 1 / \pi_{iJ}$ is created for each person interviewed, so that s_J represents the finite population U_J.

With respect to how the selection of the individuals that are dropped is made, for example in 2008, according to NSF (2012), the subsample $s_{08} \setminus s_{08(08)}$ was selected by stratifying s_{06} "into 150 strata based on three variables: demographic group, degree field, and sex." They go on to explain that:

- the past practice of selecting the sample with probability proportional to size continued, where the measure of size was the base weight associated with the previous survey cycle. For each stratum, the sampling algorithm started by identifying and removing self-representing cases through an iterative procedure. Next, the non-self-representing cases within each stratum were sorted by citizenship, disability status, degree field, and year of doctoral degree award. Finally, the balance of the sample (*i.e.*, the total allocation minus the number of self-representing cases) was selected from each stratum systematically with probability proportional to size.

It is worth mentioning that up to 1989 the cohort (or more specifically the graduation year) was part of the stratifying variables (and weight-adjustment cells), but beginning in 1991 it has not been; it was replaced by the disability status. For more details about the subsampling procedure, including the description of the sample allocation, see NSF (2012) or Cox, Grigorian, Wang and Harter (2010).

From the preceding description, it is clear that the design of the SDR is not a rotating panel design. Beside the fact that the composition of the finite population of interest is changing over time, a rotating panel design would select, at time j, a new cohort from U_j, and not from $U_j \setminus U_{j-1}$ as the SDR does.

Another peculiarity of the SDR is that, at each wave j, a frame of the recent graduates $U_{j(j)}$ exists, from which the new cohort $s_{j(j)}$ can be selected straightforwardly. However, in other applications, the cost of building such a frame, *i.e.*, a frame of new members, may be excessive (particularly as it cumulates over waves), and the new cohort may need to be selected from U_j (as opposed to from $U_{j(j)}$). The method proposed in this paper can also be applied in such cases, as long as for the total sample at wave j, s_j, a cross-sectional weight can be created to represent U_j. We further discuss this topic in Section 3.2.

Notice that in the notation $s_{j(j')}$, the quantity j represents the wave to which the sample refers, and j' denotes the sample's cohort, *i.e.*, the wave at which the sample was first selected. The notation for the weights is w_{ij}, where the first subscript identifies the subject, and the second refers to the wave of interest, regardless of when the subject was first selected.

3 Methodology

3.1 Motivation

Assume that (in a non-survey context) interest lies in the $p \times 1$ vector parameter $\boldsymbol{\beta}$ in the following model:

$$\xi : \begin{cases} E[Y_{ij} \mid X_{ij}] = \mu_{ij} = g^{-1}(X'_{ij}\boldsymbol{\beta}), & j = 1, 2, \ldots, J, i = 1, 2, \ldots \\ \text{Var}[Y_{ij} \mid X_{ij}] = \phi\nu(\mu_{ij}), & j = 1, 2, \ldots, J, i = 1, 2, \ldots \\ \text{Cov}[Y_i \mid X_i] = \Sigma_i, & i = 1, 2, \ldots \\ Y_k \perp Y_l \mid X_k, X_l, & k \neq l = 1, 2, \ldots; \end{cases} \tag{3.1}$$

where Y_{ij} is the response variable for subject i at wave j, X_{ij} is a $p \times 1$ vector of covariates, $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iJ})'$, $X_i = (X_{i1}, X_{i2}, \ldots, X_{iJ})$ is a $p \times J$ matrix; $g(\cdot)$ is a monotonic one-to-one differentiable "link function"; $\nu(\cdot)$ is the "variance function" with known form; and $\phi > 0$ is the "dispersion parameter." Since, in general, the $J \times J$ covariance matrix Σ_i is hard to specify, we model it as $\text{Cov}[Y_i \mid X_i] = V_i = A_i^{1/2}\mathbf{R}(\alpha)A_i^{1/2}$, a "working" covariance matrix; where $A_i = \text{diag}[\phi\nu(\mu_{i1}), \phi\nu(\mu_{i2}), \ldots, \phi\nu(\mu_{iJ})]$ and $\mathbf{R}(\alpha)$ is a "working" correlation matrix, both of dimension $J \times J$, and α is a vector that fully characterizes $\mathbf{R}(\alpha)$ (see Liang and Zeger 1986).

To estimate $\boldsymbol{\beta}$ we select a (single-cohort) sample of n elements from model ξ and we (intend to) measure each of them at J occasions. If all the elements in the sample respond at every single occasion j, the task can be completed with the usual generalized estimating equation (GEE) methodology of Liang and Zeger (1986). However, in any study it is rarely the case that all subjects do respond at all waves. It is more common to have some elements in the sample who drop out of the study.

Under this situation, and assuming that the missing responses can be regarded as missing at random or MAR (see Rubin 1976), in particular that the dropout at a given wave does not depend on the current (unobserved) value, Robins, Rotnitzky and Zhao (1995) proposed to estimate $\boldsymbol{\beta}$ by solving the estimating equations: $\sum_{i=1}^{n}(\partial\boldsymbol{\mu}'_i / \partial\boldsymbol{\beta})V_i^{-1}\hat{\Delta}_i(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}$, where $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{iJ})'$, $\hat{\Delta}_i = \text{diag}[R_{i1}\hat{q}_{i1}^{-1}, R_{i2}\hat{q}_{i2}^{-1}, \ldots, R_{iJ}\hat{q}_{iJ}^{-1}]$, R_{ij} is the response indicator for subject i at wave j, and $\hat{q}_{ij}$ is an estimate of the probability that subject i is observed through wave j.

For survey applications, one would use the estimating equation $\sum_{i \in s}[w_i(\partial\boldsymbol{\mu}'_i / \partial\boldsymbol{\beta})V_i^{-1}\hat{\Delta}_i(\boldsymbol{y}_i - \boldsymbol{\mu}_i)] = \mathbf{0}$, where w_i is the survey weight for subject i. Another way of writing this

equation is $\sum_{i \in s} (\partial \boldsymbol{\mu}_i' / \partial \boldsymbol{\beta}) V_i^{-1} \hat{\Delta}_{wi} (\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}$, with $\hat{\Delta}_{wi} = \text{diag}[w_i R_{i1} \hat{q}_{i1}^{-1}, w_i R_{i2} \hat{q}_{i2}^{-1}, \ldots, w_i R_{iJ} \hat{q}_{iJ}^{-1}]$.

We notice that the diagonal elements of $\hat{\Delta}_{wi}$ are simply wave-specific nonresponse-adjusted survey weights whenever the subject is observed, and are equal to zero whenever the subject is missing. This feature in and of itself suggests a solution to the multi-cohort problem, which will be presented in the next section.

3.2 A novel approach to combining cohorts in longitudinal surveys

Based on the discussion in the previous section, if we have a fixed-panel, fixed-panel-plus-'births', repeated-panel, rotating-panel, split-panel, or refreshment sample survey, we propose to estimate the superpopulation parameter $\boldsymbol{\beta}$ in model ξ by the solution to the estimating equations:

$$\Psi_s(\boldsymbol{\beta}) = \sum_{i \in s} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} W_i (\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}; \tag{3.2}$$

where the sum is over the sample s, *i.e.*, over all the elements selected (for the first time) in any of the samples $s_{1(1)}, s_{2(2)}, \ldots, s_{J(J)}$. The diagonal matrix W_i is $W_i = \text{diag}[I_i(U_1) w_{i1}, I_i(U_2) w_{i2}, \ldots, I_i(U_J) w_{iJ}]$, with w_{ij} being the (nonresponse-adjusted) cross-sectional weight for subject i at wave j (as long as subject i is part of sample s_j) and $I_i(U_j)$ is the indicator of whether subject i belongs to finite population U_j or not. In Section 3.2.1 we argue why this is a reasonable estimation procedure, and in Section 3.2.2 we discuss the missing value issue.

The cross-sectional weights w_{ij}, in W_i, are such that the sample s_j represents U_j, when used in conjunction with said weights. This means that, for each observation i in sample s_j, there has to be a survey weight w_{ij}, which could be regarded as the number of units that such observation represents in U_j. However, remember that the sample s_j is composed of different sets of subjects, or different subsamples (the different cohorts), and the integration of these subsamples into a single cross-sectional weight variable w_{ij} may not be a straightforward task.

For the SDR, the construction of the cross-sectional weight for wave j is not too complicated as the different cohorts are selected independently, from non-overlapping populations. The base weight in that case is easy to compute, and all that remains is the adjustment for things like attrition and calibration to known totals in the population U_j.

On the other hand, in other situations, for example, when a frame of *new* members does not exist, the new cohort may need to be selected from the overall population at the given wave, or from a frame containing new members *plus* some old members, or from multiple frames. In such cases, the building of the cross-sectional weights may not be as straightforward, and the theory of multiple frames may need to be used. We refer the reader to the works of Lohr (2007) and Rao and Wu (2010), and references therein, for cases like that.

Expression (3.2) is a generalization of equation (2.25) in Vieira (2009). The latter is applicable only when all the subjects have the same number of observations or any missing responses can be regarded as missing completely at random or MCAR (see Rubin 1976). As discussed in Robins, *et al.* (1995), using such an equation when the missing responses are not MCAR produces inconsistent estimators; therefore, with a rotation scheme like that of the SDR, where not all subjects are dropped (or kept) with the same probabilities, its usage would not be appropriate. The adequacy of equation (3.2) in that case and when there are missing responses is addressed in sections 3.2.1 and 3.2.2, respectively. If all subjects have cross-sectional weights that do not vary over time (or have a single longitudinal weight) equation (3.2) reduces to equation (2.25) in Vieira (2009).

3.2.1 Unbiasedness

The unbiasedness property of the estimating function is important because, as Song (2007, Section 5.4) argues, it is the most crucial assumption in order to obtain a consistent estimator.

Let us define β_N, the so-called "census estimator," to be the solution to the following finite population estimating equation:

$$\Psi_U(\boldsymbol{\beta}_N) = \sum_{i \in U} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}_N} V_i^{-1} \mathrm{I}_i(\mathrm{U})(\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}_N)) = \mathbf{0}, \tag{3.3}$$

where the sum is over U, *i.e.*, over all the elements who became members of the target population in any of $U_{1(1)}, U_{2(2)}, \ldots, U_{J(J)}$, and $\mathrm{I}_i(\mathrm{U}) = \mathrm{diag}[I_i(U_1), I_i(U_2), \ldots I_i(U_J)]$. In order to show design-unbiasedness of the estimating function $\Psi_s(\boldsymbol{\beta})$, we need to show that its design expectation is $\Psi_U(\boldsymbol{\beta})$ for any $\boldsymbol{\beta}$.

The sampling design characteristics of a longitudinal survey can be thought of as those of a multiphase sample, as can be seen in Särndal, Swensson and Wretman (1992, Section 9.9). We therefore use the methodology of multiphase sampling for the derivations. We assume, without loss of generality, that there are only three waves; the derivations with just three waves show the patterns for general J, with respect to unbiasedness and variance.

As we mentioned earlier, we assume that w_{ij} is the cross-sectional weight for subject i at wave j, if that subject belongs to s_j, and zero otherwise. From the theory of multiphase sampling we have that for $i \in s_{1(1)}$, $w_{i1} = \pi_{i1}^{-1}$, $w_{i2} = \pi_{i1}^{-1}\pi_{i2|s_{1(1)}}^{-1}$, and $w_{i3} = \pi_{i1}^{-1}\pi_{i2|s_{1(1)}}^{-1}\pi_{i3|s_{2(1)}}^{-1}$; for $i \in s_{2(2)}$, $w_{i2} = \pi_{i2}^{-1}$ and $w_{i3} = \pi_{i2}^{-1}\pi_{i3|s_{2(2)}}^{-1}$; and for $i \in s_{3(3)}$, $w_{i3} = \pi_{i3}^{-1}$; where π_{ij} is the inclusion probability of subject i in sample $s_{j(j)}$ and $\pi_{ij|s_{j-1(j')}}$ is the conditional inclusion probability of subject i in sample $s_{j(j')}$ given $s_{j-1(j')}$.

Using $E_p(\cdot)$ to denote the expectation with respect to the sampling design, we have:

$$E_p\left[\sum_{i \in s} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} W_i(\boldsymbol{y}_i - \boldsymbol{\mu}_i)\right] = E_p\left[\sum_{j=1}^{3} \sum_{i \in s_{j(j)}} B_i W_i \boldsymbol{e}_i\right]; \tag{3.4}$$

where $B_i = (\partial \boldsymbol{\mu}_i' / \partial \boldsymbol{\beta}) V_i^{-1}$ and $\boldsymbol{e}_i = \boldsymbol{y}_i - \boldsymbol{\mu}_i$. For example, for $\sum_{i \in s_{2(2)}} B_i W_i \boldsymbol{e}_i$ we obtain:

$$\begin{aligned} E_p\left[\sum_{i \in s_{2(2)}} B_i W_i \boldsymbol{e}_i\right] &= E\left\{E\left[\sum_{i \in U_{2(2)}} B_i D_i \boldsymbol{e}_i \mid s_{2(2)}\right]\right\} = E\left\{\sum_{i \in U_{2(2)}} B_i D_i^{*} \boldsymbol{e}_i\right\} \\ &= \sum_{i \in U_{2(2)}} B_i D_i^{**} \boldsymbol{e}_i \overset{\text{def}}{=} \sum_{i \in U_{2(2)}} B_i \mathrm{I}_i(\mathrm{U}) \boldsymbol{e}_i, \end{aligned} \tag{3.5}$$

where $D_i = \text{diag}[0, I_i(U_2)w_{i2}I_i(s_{2(2)}), I_i(U_3)w_{i3}I_i(s_{3(2)})I_i(s_{2(2)})]$, $D_i^* = \text{diag}[0, (I_i(U_2)w_{i2}I_i(s_{2(2)})), (I_i(U_3)\pi_{i3|s_{2(2)}}I_i(s_{2(2)})) / (\pi_{i2}\pi_{i3|s_{2(2)}})]$, and $D_i^{**} = \text{diag}[0, (I_i(U_2)\pi_{i2}) / \pi_{i2}, \ (I_i(U_3)\pi_{i2}) / \pi_{i2}]$; similarly we can show that $E_p\left[\sum_{i\in s_{1(1)}} B_i W_i \boldsymbol{e}_i\right] = \sum_{i\in U_{1(1)}} B_i \mathrm{I}_i(\mathrm{U})\boldsymbol{e}_i$ and $E_p\left[\sum_{i\in s_{3(3)}} B_i W_i \boldsymbol{e}_i\right] = \sum_{i\in U_{3(3)}} B_i \mathrm{I}_i(\mathrm{U})\boldsymbol{e}_i$. From these expressions and equation (3.4) we conclude that $E_p[\Psi_s(\boldsymbol{\beta})] = \Psi_U(\boldsymbol{\beta})$ for any $\boldsymbol{\beta}$, which means that the estimating function $\Psi_s(\boldsymbol{\beta})$ is design-unbiased for the finite population estimating function.

Furthermore, as the target of inference is the superpopulation parameter, we need to guarantee that the model for μ_{ij} is such that $E_\xi(Y_{ij} - \mu_{ij}) = 0$ is satisfied, where $E_\xi(\cdot)$ represents the expectation with respect to model ξ. For if this is the case, we have:

$$E_{\xi p}[\Psi_s(\boldsymbol{\beta})] \overset{\text{def}}{=} E_\xi E_p[\Psi_s(\boldsymbol{\beta})] = E_\xi[\Psi_U(\boldsymbol{\beta})] = \sum_{i\in U} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} V_i^{-1} \mathrm{I}_i(\mathrm{U}) E_\xi(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \mathbf{0};$$

so that the estimating function $\Psi_s(\boldsymbol{\beta})$ is model-design unbiased. The requirement $E_\xi(Y_{ij} - \mu_{ij}) = 0$ means that the mean model needs to be correctly specified; consequently, one needs to pay attention to residual diagnostics for the particular model being fitted.

3.2.2 A note on nonresponse

In the SDR, as in any other (longitudinal) survey, there is nonresponse. Some sampled individuals choose not to participate at all, whereas some subjects participate in some waves but not in others. The SDR remedies this situation by making a nonresponse adjustment to the cross-sectional survey weights.

Assume that the nonresponse adjustment at wave j is a multiplication by the inverse of the estimated wave j response probability $\hat{\pi}_{rij}$. For example, the nonresponse-adjusted weight for a person who *did* respond at wave 3 (and was first selected at wave 2), *i.e.*, for $i \in r_{3(2)}$, would be $w_{ri3} = \pi_{i2}^{-1}\pi_{i3|s_{2(2)}}^{-1}\hat{\pi}_{ri3}^{-1}$.

We need to redefine the estimating equation, to include only the respondents, as $\Psi_r(\boldsymbol{\beta}) = \sum_{i\in r}(\partial \boldsymbol{\mu}_i' / \partial \boldsymbol{\beta}) V_i^{-1} W_{ri}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}$, where the sum is over the respondent set r, *i.e.*,

over all the elements who belonged for the first time in any of the respondent sets $r_{1(1)}, r_{2(2)}, \ldots, r_{J(J)}$, and the matrix W_{ri} is $W_{ri} = \text{diag}[I_i(U_1)w_{ri1}, I_i(U_2)w_{ri2}, \ldots, I_i(U_J)w_{riJ}]$. Also, denote by $r_{j(j')}$ the set of cohort j' respondents at wave j. Obviously, $w_{rij} = 0$ if $i \notin r_j = \bigcup_{j'=1}^{j} r_{j(j')}$.

If additionally, the response mechanism (R) can be assumed to be MAR, we then have, for example for $\sum_{i \in r_{2(2)}} B_i W_{ri} \boldsymbol{e}_i$:

$$E_R\left\{\sum_{i\in r_{2(2)}} B_i W_{ri}\boldsymbol{e}_i\right\} = E_R\left\{\sum_{i\in s_{2(2)}} B_i D_i \boldsymbol{e}_i\right\} = \sum_{i\in s_{2(2)}} B_i D_i^{*}\boldsymbol{e}_i = \sum_{i\in s_{2(2)}} B_i D_i^{**}\boldsymbol{e}_i \overset{\text{def}}{=} \sum_{i\in s_{2(2)}} B_i W_i \boldsymbol{e}_i, \quad (3.6)$$

where $D_i = \text{diag}[0, I_i(U_2)w_{ri2}I_i(r_{2(2)}), I_i(U_3)w_{ri3}I_i(r_{3(2)})]$, $D_i^{*} = \text{diag}[0, (I_i(U_2)\pi_{ri2}) / (\pi_{i2} \times \hat{\pi}_{ri2}), (I_i(U_3)\pi_{ri3}) / (\pi_{i2}\pi_{i3|s_{2(2)}}\hat{\pi}_{ri3})]$, and $D_i^{**} = \text{diag}[0, I_i(U_2)w_{i2}, I_i(U_3)w_{i3}]$. The third equality in (3.6) requires that the nonresponse model used for $\hat{\pi}_{rij}$ satisfies $E_R[I_i(r_{j(j')})] \overset{\text{def}}{=} \pi_{rij} = \hat{\pi}_{rij}$. This means that in the model for $\hat{\pi}_{rij}$ we have to include as much available information, thought to influence the nonresponse propensity, as possible, in order for this assumption (*i.e.*, the MAR assumption) to be tenable. For example, if the nonresponse is thought to be independent across waves, one should include, in the model for $\hat{\pi}_{rij}$, as many variables from the corresponding wave as possible. If, on the other hand, it is reasonable to assume that the response propensity at a given wave depends on previous responses (and possibly response history), then those responses should be included in the response model, and so on.

The design as well as the model-design unbiasedness follow immediately from (3.6) together with the previous section. Hereafter we therefore ignore the issue of nonresponse for notational simplicity.

3.3 Variance and variance estimation

We now develop a (Taylor Series) linearization for the variance of the proposed estimator. The basic technique is due to Binder (1983). For simplicity in the derivations and notation we divide through by N; we redefine

$$\Psi_s(\boldsymbol{\beta}) = N^{-1}\sum_{i\in s}\frac{\partial\boldsymbol{\mu}_i'}{\partial\boldsymbol{\beta}}V_i^{-1}W_i(\boldsymbol{y}_i - \boldsymbol{\mu}_i) \text{ and } \Psi_U(\boldsymbol{\beta}) = N^{-1}\sum_{i\in U}\frac{\partial\boldsymbol{\mu}_i'}{\partial\boldsymbol{\beta}}V_i^{-1}\mathrm{I}_i(\mathrm{U})(\boldsymbol{y}_i - \boldsymbol{\mu}_i),$$

where $N = \sum_{j=1}^{J} N_j$. Let $\hat{\boldsymbol{\beta}}$ be our estimator, which satisfies $\Psi_s(\hat{\boldsymbol{\beta}}) = \mathbf{0}$, and let $\boldsymbol{\beta}_N$ be the "census estimator," which satisfies $\Psi_U(\boldsymbol{\beta}_N) = \mathbf{0}$. Assume $\boldsymbol{\beta}_N - \boldsymbol{\beta} = O_P(1/\sqrt{N_m})$ and $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N = O_P(1/\sqrt{n_m})$, with $N_m = \min\{N_1, N_2, \ldots, N_J\}$ and $n_m = \min\{n_1, n_2, \ldots, n_J\}$. We can write the total error of $\hat{\boldsymbol{\beta}}$ as $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N) + (\boldsymbol{\beta}_N - \boldsymbol{\beta})$ = Sampling Error + Model Error. After some straightforward calculations, the total variance, or more precisely the total MSE, can be decomposed as:

$$V_{\text{Tot}} = E_{\xi p}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' = V_{\text{Sam}} + 2 \otimes C_{\text{Sam-Mod}} + o(1/n_m), \tag{3.7}$$

where $2 \otimes A = A + A'$ for any matrix A, $V_{\text{Sam}} = E_\xi V_p$ is the "sampling variance" component, $2 \otimes C_{\text{Sam-Mod}}$ is the cross "sampling-model variance" component, $V_p = E_p[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N)']$, $C_{\text{Sam-Mod}} = E_p C_\xi$, and $C_\xi = E_\xi(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\boldsymbol{\beta}_N - \boldsymbol{\beta})'$. Furthermore, by Taylor series expansions we can obtain the following approximations: $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N = [H(\boldsymbol{\beta}_N)]^{-1}\Psi_s(\boldsymbol{\beta}_N) + o_P(1/\sqrt{n_m})$, $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = [\hat{H}(\boldsymbol{\beta})]^{-1}\Psi_s(\boldsymbol{\beta}) + o_P(1/\sqrt{n_m})$, and $\boldsymbol{\beta}_N - \boldsymbol{\beta} = [H(\boldsymbol{\beta})]^{-1}\Psi_U(\boldsymbol{\beta}) + o_P(1/\sqrt{N_m})$, where we define $H(\boldsymbol{\beta}) = N^{-1}\sum_{i\in U}(\partial\boldsymbol{\mu}_i'/\partial\boldsymbol{\beta})V_i^{-1}\mathrm{I}_i(\mathrm{U})(\partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta})$ and $\hat{H}(\boldsymbol{\beta}) = N^{-1}\sum_{i\in s}(\partial\boldsymbol{\mu}_i'/\partial\boldsymbol{\beta})V_i^{-1}W_i(\partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta})$.

We then get, for V_p and C_ξ in (3.7),

$$V_p = [H(\boldsymbol{\beta}_N)]^{-1}\mathrm{Var}_p[\Psi_s(\boldsymbol{\beta}_N)][H(\boldsymbol{\beta}_N)]^{-1} + o_P(1/n_m), \tag{3.8}$$

$$\begin{aligned} C_\xi &= [\hat{H}(\boldsymbol{\beta})]^{-1}E_\xi[\Psi_s(\boldsymbol{\beta})\Psi_U'(\boldsymbol{\beta})][H(\boldsymbol{\beta})]^{-1} + o_P(1/n_m) \\ &= N^{-1}[\hat{H}(\boldsymbol{\beta})]^{-1}\hat{H}_{\Sigma V}(\boldsymbol{\beta})[H(\boldsymbol{\beta})]^{-1} + o_P(1/n_m), \end{aligned} \tag{3.9}$$

where $\mathrm{Var}_p[\Psi_s(\boldsymbol{\beta}_N)] = E_p[\Psi_s(\boldsymbol{\beta}_N)\Psi_s'(\boldsymbol{\beta}_N)]$ and $\hat{H}_{\Sigma V}(\boldsymbol{\beta}) = N^{-1}\sum_{i\in s}[(\partial\boldsymbol{\mu}_i'/\partial\boldsymbol{\beta})V_i^{-1}W_i\Sigma_i \times V_i^{-1}(\partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta})]$; the derivation of (3.9) can be found in the Appendix.

In conclusion, so far we have found that:

$$
\begin{aligned}
V_{\text{Tot}} &= E_\xi V_p + 2 \otimes E_p C_\xi + o(1/n_m) \\
&= E_\xi \left\{ [H(\boldsymbol{\beta}_N)]^{-1} \text{Var}_p[\Psi_s(\boldsymbol{\beta}_N)][H(\boldsymbol{\beta}_N)]^{-1} \right\} \\
&\quad + 2 \otimes N^{-1} E_p \left\{ [\hat{H}(\boldsymbol{\beta})]^{-1} \hat{H}_{\Sigma V}(\boldsymbol{\beta})[H(\boldsymbol{\beta})]^{-1} \right\} + o(1/n_m).
\end{aligned} \tag{3.10}
$$

In (3.10) all the terms can be estimated by "plugging in" the estimate $\hat{\boldsymbol{\beta}}$, except for the term $\text{Var}_p[\Psi_s(\boldsymbol{\beta}_N)]$; this is the subject of the next section.

If the sampling fraction is small, *i.e.*, $n \ll N$, the first term in expression (3.10) is a good approximation for the total variance; *i.e.*, the expression for V_{Tot} is simply $E_\xi V_p$ (and lower order terms). If, on the other hand, the sampling fraction is large, both terms in (3.10) are required.

3.3.1 Design variance of the estimating function

In order to derive an expression for $\text{Var}_p[\Psi_s(\boldsymbol{\beta}_N)]$, we assume $J = 3$, as before. The methodology is that of two-phase sampling (more precisely, multiphase sampling), as discussed in chapter 9 of Särndal, *et al.* (1992). After some derivations (see Appendix), and defining $B_i = (\partial \boldsymbol{\mu}_i' / \partial \boldsymbol{\beta})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_N} V_i^{-1}$, $\boldsymbol{e}_i = \boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}_N)$, $\boldsymbol{e}_{i(1\cdots3)} = \boldsymbol{e}_i$, $\boldsymbol{e}_{i(2\cdots3)} = (0, e_{i2}, e_{i3})'$, and $\boldsymbol{e}_{i(3\cdots3)} = (0, 0, e_{i3})'$, we obtain:

$$
\text{Var}_p[\Psi_s(\boldsymbol{\beta}_N)] = \sum_{j=1}^{3} D_{(j)} = \sum_{j=1}^{3}\sum_{k=j}^{3} D_{(j)k}, \tag{3.11}
$$

where $D_{(j)} \overset{\text{def}}{=} N^{-2}\text{Var}_p\left(\sum_{i \in s_{j(j)}} B_i W_i \boldsymbol{e}_i \right) = \sum_{k=j}^{3} D_{(j)k}$, for $j = 1, 2, 3$,

$$
N^2 D_{(j)j} \overset{\text{def}}{=} \text{Var}\left[\sum_{i \in s_{j(j)}} w_{ij} B_i \text{I}_i(\text{U}) \boldsymbol{e}_{i(j\cdots3)} \right], \text{ for } j = 1, 2, 3,
$$

$$
N^2 D_{(j-1)j} \overset{\text{def}}{=} E\left\{ \text{Var}\left[\sum_{i \in s_{j(j-1)}} w_{ij} B_i \text{I}_i(\text{U}) \boldsymbol{e}_{i(j\cdots3)} \mid s_{j-1(j-1)} \right] \right\}, \text{ for } j = 2, 3,
$$

$$
N^2 D_{(1)3} \overset{\text{def}}{=} E\left\{ E\left[\text{Var}\left(\sum_{i \in s_{3(1)}} w_{i3} B_i \text{I}_i(\text{U}) \boldsymbol{e}_{i(3\cdots3)} \mid s_{2(1)}, s_{1(1)} \right) \mid s_{1(1)} \right] \right\},
$$

and in the Appendix we show that:

$$N^2 D_{(j)k} = \mathrm{Var}\left[\sum_{i\in s_{k(j)}} w_{ik} B_i \mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(k\cdots 3)}\right] - \mathrm{Var}\left[\sum_{i\in s_{k-1(j)}} w_{i,k-1} B_i \mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(k\cdots 3)}\right],$$

for $j = 1, 2, 3,$ and $3 \geq k > j$. In general, we have proved the following

Property 3.1 The (design) variance of $\Psi_s(\boldsymbol{\beta}_N)$ can be decomposed as:

$$\mathrm{Var}_p[\Psi_s(\boldsymbol{\beta}_N)]$$

$$= \frac{1}{N^2}\sum_{j'=1}^{J}\sum_{j=j'}^{J}\left\{\mathrm{Var}_p\left[\sum_{i\in s_{j(j')}} w_{ij} B_i \mathrm{I}_i(U)\mathbf{e}_{i(j\cdots J)}\right] - \mathrm{Var}_p\left[\sum_{i\in s_{j-1(j')}} w_{i,j-1} B_i \mathrm{I}_i(U)\mathbf{e}_{i(j\cdots J)}\right]\right\} \quad (3.12)$$

$$= \frac{1}{N^2}\sum_{j=1}^{J}\left\{\mathrm{Var}_p\left[\sum_{i\in s_j} w_{ij} B_i \mathrm{I}_i(U)\mathbf{e}_{i(j\cdots J)}\right] - \mathrm{Var}_p\left[\sum_{i\in s_{j-1}} w_{i,j-1} B_i \mathrm{I}_i(U)\mathbf{e}_{i(j\cdots J)}\right]\right\}, \quad (3.13)$$

where we let $w_{i,j-1} = 0$ whenever $j = j'$, $w_{i0} = 0,$ and to get (3.13) we have changed variables and used the independence among cohorts.

In (3.11), (3.12), and (3.13) we have assumed that the cohorts are design-independent. However, in some cases this assumption may not be tenable; an example of such a case is the multiple frame situation discussed in the first part of Section 3.2. Another instance in which it may not be appropriate to assume cohort independence is when weight adjustments cross cohorts, which is the case of the SDR; we discuss this issue in Section 5. Calculations for the case of three cohorts, in the Appendix, show that (3.13) holds for the variance terms even without independence. The Appendix also identifies conditions under which it is a good approximation for the covariance terms.

3.3.2 Estimation

The estimation of V_{Tot} in (3.10) can be achieved as follows. $H(\boldsymbol{\beta}_N)$, $\hat{H}(\boldsymbol{\beta})$, and $H(\boldsymbol{\beta})$ can be estimated by $\hat{H}(\hat{\boldsymbol{\beta}})$. $\hat{H}_{\Sigma V}(\boldsymbol{\beta})$ can be estimated by $\hat{H}_{\Sigma V}(\hat{\boldsymbol{\beta}})$, where $\Sigma_i = \mathrm{Cov}[Y_i \mid X_i]$ can be estimated by $\hat{\boldsymbol{e}}_i\hat{\boldsymbol{e}}_i'$.

We use (3.13) in Property 3.1 to estimate $\mathrm{Var}_p[\Psi_s(\boldsymbol{\beta}_N)]$. As long as there is a method to estimate the variance of (cross-sectional) Horvitz-Thompson (H-T) estimators, expression (3.13)

can be used. If we define $Z_{ij} = B_i \mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(j\cdots J)}$, we notice that each of the terms involved in the computation of (3.13), terms like $\mathrm{Var}_p\left[\sum_{i\in s_j} w_{ij} Z_{ij}\right]$, is simply the variance of a wave-j H-T estimator. Obviously, the variance estimation method needs to account for the sampling design as well as for any nonresponse and calibration adjustments performed, but this does not present any additional complications beyond what is found in any cross-sectional problem, as everything is implemented cross-sectionally. The SDR uses replication to estimate variances of cross-sectional estimators, but any method of design variance estimation can be used.

We use the cross-sectional replicate weights that SDR provides, but we do not re-estimate the parameter of interest at each replicate. First, note that we require replication only for the estimation of the "meat" $(\mathrm{Var}_p[\Psi_s(\boldsymbol{\beta}_N)])$ of the design variance $(E_\xi V_p)$. Secondly, although $\hat{\boldsymbol{\beta}}$ does appear in the expression for the H-T estimator whose variance needs to be calculated (and re-calculated at each replicate), the work of Roberts, Binder, Kovačević, Pantel and Phillips (2003), who apply the "estimating function bootstrap" (Hu and Kalbfleisch 2000) to survey data, show that in a setting like ours, it is not necessary to re-compute the estimator at each replicate, but that the full-sample estimator suffices. This simplification speeds up the computation of the replicate estimates.

As a way of illustration, say we currently are at wave j, *i.e.*, we are estimating the j^{th} term in (3.13). The r^{th} replicate of the first term is $\sum_{i\in s_j} w_{ij}^{(r)} B_i(\hat{\boldsymbol{\beta}})\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(j\cdots J)}(\hat{\boldsymbol{\beta}})$, where $w_{ij}^{(r)}$ is the r^{th} replicate weight for subject i at wave j, and the r^{th} replicate of the second term is $\sum_{i\in s_{j-1}} w_{i,j-1}^{(r)} B_i(\hat{\boldsymbol{\beta}})\mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(j\cdots J)}(\hat{\boldsymbol{\beta}})$, where $w_{i,j-1}^{(r)}$ is the r^{th} replicate weight for subject i at wave $j-1$.

4 Application to the SDR

The dataset we use is the restricted SDR data, under a license agreement from NSF. The SDR collects information about employment situation, principal employer, principal job, past employment, recent education, demographics, and disability, among others that vary from wave

to wave. We use only information requested in all the waves of interest: 1995, 1997, 1999, 2001, 2003, 2006, and 2008.

To illustrate our methodology, we constructed a model for individuals' salaries over time. The response is the log of salary (in the principal job), with an identity link function, and several covariates; modeling log of salary (as opposed to salary) is a standard practice. There are both time-independent covariates (such as gender) and time-dependent ones (such as employment sector). We have four major classes of covariates. The *Degree variables* are: degree field, years since degree, and age at graduation. The *Job variables* are: job field or category, sector, postdoc indicator, adjunct faculty indicator, hours worked per week in the principal job, weeks per year in the principal job, how related is the job to the doctoral degree, part-time for different reasons, number of months since started in the principal job, the starting month in the principal job, whether the employer/type of job has changed since previous wave, and whether changed employer/type of job since previous wave because was laid off or job terminated. The *Person's demographics* are: gender, citizenship status, race/ethnicity, presence of children in family, marital status, and spouse's working status. Finally, the "*Environment*" *variables* are: years since 1995, state (of employment), and the consumer price index (of the region of employment). The full list of variables, interactions, and categories can be found in Carrillo and Karr (2011). For categorical variables, the reference category is the one with the largest count.

The dataset for our model consists of 59,346 subjects and 190,693 observations, distributed as: $n_{95} = 30{,}234$, $n_{97} = 30{,}652$, $n_{99} = 26{,}732$, $n_{01} = 26{,}778$, $n_{03} = 24{,}956$, $n_{06} = 25{,}910$, and $n_{08} = 25{,}431$. Those data correspond to non-missing salaries between \$5,000 and \$999,995, for people with consistent ages across the waves, and with non-missing value for the variable indicating whether the (postsecondary educational institution) employer was public or private. The average (cross-sectional) survey weight for each of those waves are: $\bar{w}_{95} = 15.37$, $\bar{w}_{97} = 16.28$, $\bar{w}_{99} = 19.96$, $\bar{w}_{01} = 20.74$, $\bar{w}_{03} = 22.71$, $\bar{w}_{06} = 22.93$, and $\bar{w}_{08} = 24.88$.

The survey weights that we use for each wave are the final adjusted weights. These weights are the original design weights adjusted for nonresponse and post-stratification. However, the theory that we developed in Section 3 assumes that the weights are the inverse of the selection probabilities; in other words, the original design weights. This is a mismatch whose effect we plan to investigate in the future. On the other hand, the calculations in the last part of the Appendix (which do not assume anything about the weights) suggest that the effect of this mismatch is small.

The covariates and interactions that we considered were selected because they were suggested either by exploratory analyses or by the subject matter experts at the NSF. Carrillo and Karr (2011) present the estimated β coefficients in the model $y_{ij} = \log(\text{SALARY}_{ij}) = X'_{ij}\boldsymbol{\beta} + \varepsilon_{ij}$, where X_{ij} includes the intercept along with the other covariates. This $\boldsymbol{\beta}$ corresponds to the one in model ξ, in Formula (3.1), and whose properties are discussed in Section 3. The working covariance matrix is estimated to be $\hat{V}_i = \hat{\phi}\mathbf{R}(\hat{\alpha})$, with $\hat{\phi} = \hat{\sigma}^2 = \left(\sum_{i\in s}\sum_{j=95}^{08} w_{ij}\hat{e}_{ij}^2\right)\Big/\left(\sum_{i\in s}\sum_{j=95}^{08} w_{ij} - p\right) = 0.196$, where $\hat{e}_{ij} = y_{ij} - X'_{ij}\hat{\boldsymbol{\beta}}$ and $p = 208$ is the number of covariates in X_{ij}, w_{ij} is the cross-sectional weight for subject i at wave j as long as $i \in s_j$ and zero otherwise. The estimate $\hat{\alpha}$ contains the $21 = (7 \times 6) / 2$ estimated auto-correlations $\hat{\alpha}_{jj'} = \hat{\alpha}_{j'j} = \left(\sum_{i\in s}\sqrt{w_{ij}}\sqrt{w_{ij'}}\hat{e}_{ij}\hat{e}_{ij'}\right)\Big/\left(\hat{\phi}\left[\sum_{i\in s}\sqrt{w_{ij}}\sqrt{w_{ij'}} - p\right]\right)$, for $j \neq j' =$ 1995, 1997, 1999, 2001, 2003, 2006, 2008, and $\hat{\alpha}_{jj} = 1$ for all j. These estimated values form the auto-correlation matrix:

$$\mathbf{R}(\hat{\alpha}) = \begin{pmatrix} 1 & \hat{\alpha}_{95,97} & \hat{\alpha}_{95,99} & \hat{\alpha}_{95,01} & \hat{\alpha}_{95,03} & \hat{\alpha}_{95,06} & \hat{\alpha}_{95,08} \\ & 1 & \hat{\alpha}_{97,99} & \hat{\alpha}_{97,01} & \hat{\alpha}_{97,03} & \hat{\alpha}_{97,06} & \hat{\alpha}_{97,08} \\ & & 1 & \hat{\alpha}_{99,01} & \hat{\alpha}_{99,03} & \hat{\alpha}_{99,06} & \hat{\alpha}_{99,08} \\ & & & 1 & \hat{\alpha}_{01,03} & \hat{\alpha}_{01,06} & \hat{\alpha}_{01,08} \\ & & & & 1 & \hat{\alpha}_{03,06} & \hat{\alpha}_{03,08} \\ & \text{sym} & & & & 1 & \hat{\alpha}_{06,08} \\ & & & & & & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0.38 & 0.36 & 0.32 & 0.30 & 0.28 & 0.27 \\ & 1 & 0.42 & 0.36 & 0.33 & 0.32 & 0.31 \\ & & 1 & 0.46 & 0.38 & 0.36 & 0.34 \\ & & & 1 & 0.47 & 0.40 & 0.38 \\ & & & & 1 & 0.49 & 0.44 \\ & \text{sym} & & & & 1 & 0.55 \\ & & & & & & 1 \end{pmatrix}.$$

We now give some conclusions about salaries in the Ph.D. workforce based on the estimated coefficients, which appear in Carrillo and Karr (2011). First of all, a sensible estimate of mean salary considers the intercept, the hours worked per week (whose average is 47), and years since degree (average of 15); so that an estimate of the overall average is $\exp(9.4 + 47 \times 0.038 - 47^2 \times 0.0003 + 15 \times 0.03 - 15^2 \times 0.0006) = \$52,067$, for a subject with all other continuous covariates equal to zero and in the reference of all categorical covariates.

All other things being constant, women's salaries are about 93.4% those of men, whereas race does not seem to have an effect on salaries. The gender $\times$ years since 1995 interaction is not significant; therefore this salary differential is not changing over time. Notice that with a single year's data, we would not be able to evaluate the effect of time. Even more important than that, using only the data from a single wave, say 2008, we would not be able to assess whether the effect of being female is changing over time.

Doctorate holders with a management job have the highest salaries, followed by those in health occupations; on the other hand, those with the lowest salaries are the ones employed in "other" occupations, followed by those in political science.

Among employment sectors, salaries are highest in for-profit industry (20% higher than for the reference category of tenured faculty in public 4-year institutions), followed in order by the federal government, self-employment, non-profit industry, all of which are higher than the reference category. The lowest salaries are those in two-year colleges and in two- and four-year institutions for which tenure is not applicable.

The highest single negative effect on salaries also occurs within the education sector. Those with positions as adjunct faculty members have salaries that are approximately 59% of the salaries of comparable doctorate holders. Not surprisingly, postdoctoral salaries are only about 74% of the salaries of comparable people in other types of positions.

Sector is also a contributing factor to the hard-to-interpret dependence of salary on the starting month for the current position: salaries are lower for starting months of August and September.

Additional analyses show that the monthly effect is present only in the education sector, where, as we have seen, salaries are lower than in industry or government, and in which starting months of August and September are common. Therefore, sector is part of the answer, but not the entire answer. Finer-grained divisions of the education sector, using Carnegie classifications, further reduce, but do not remove, the significance of monthly effects. The SDR does not seem to contain sufficient data to remove the monthly effects entirely, so we have retained the SDR definition of sector.

People with degrees in computing and information sciences have the highest salaries (around 20% higher than in the biological sciences), followed by those in electrical and computer engineering and in economics (approximately 16% higher). Doctorate holders in agricultural and food sciences, environmental life sciences, earth, atmospheric, and ocean sciences, and in "other" social sciences have the lowest salaries. The "other" social sciences are the social sciences excluding economics and political science.

Married people have the highest salaries, followed by those who are in married-like relationships, widowed, separated, divorced, and never married. The latter have salaries only around 89% as high as the married ones; one could argue that there is some association between never married and age. The presence of children older than two is associated with higher salaries, but the presence of children younger than two is not.

Doctorate holders with jobs only somewhat related to their degree field make around 93% of what people with closely related jobs (the reference category) do. If the job is not related to the doctoral degree as the result of a change in career or professional interests, they make around 82% of what people with closely related jobs do. On the other hand, those with jobs not related for other reasons make only about 76% of what the reference category does.

There is an increase of around 3% for every additional year since doctorate graduation, although there is a diminishing effect for higher number of years. We interpret this as the effect

of experience. There is a small penalty for receiving the doctorate later in life; for every additional year of age at graduation, the salary reduces by 1%.

We also found that the regional Consumer Price Index (CPI) is significant. The higher the CPI, the higher the salary. We could not use the CPI associated with the labor market of employment because the SDR data do not identify geography beyond the state. We included the state in the model as a proxy for cost of living; the state effect is highly significant and some state coefficients are among the highest overall. The highest salaries are in California, Washington D.C. and its suburbs, and New York City and its suburbs. On the other hand, the lowest salaries are in Puerto Rico, Vermont, Montana, Maine, Idaho, South Dakota, North Dakota, and in the Territories/Abroad.

Having a part-time job due to being retired or semi-retired is significant and in several significant interactions. Because of this, we do not think that the available data present the full picture about retirement, for example, for people who are (semi-)retired and yet have full-time jobs.

Finally, we analyzed residuals; Figures 4.1 and 4.2 show a Box and Whisker plot of standardized residuals by year and a spaghetti plot of standardized residuals, respectively.

Figure 4.1 shows that the model fits reasonably well for all the reference years as most of the standardized residuals lie between -2 and 2. Also, the distributions of residuals do not seem to greatly differ from year to year.

From Figure 4.2 we also conclude that the model fits reasonably well for most people, as most of the lines fluctuate between -2 and 2. Nonetheless, there are a few people for which the model seems to greatly over-predict in 2003 and some few people for whom that happens in 2006. We included several terms in the model to correct this issue but clearly none seemed to do so completely.

The last thing we tried was to produce exploratory classification trees for these residual blips. We found that, in the dataset available, the only thing related to them was the survey mode. The

blips in 2003 are disproportionately high for web responses, and the blips in 2006 are disproportionately high for CATI responses. We conclude that either there is a mode effect in these two years or those respondents have something different, in those years, that is not included in the available variables.

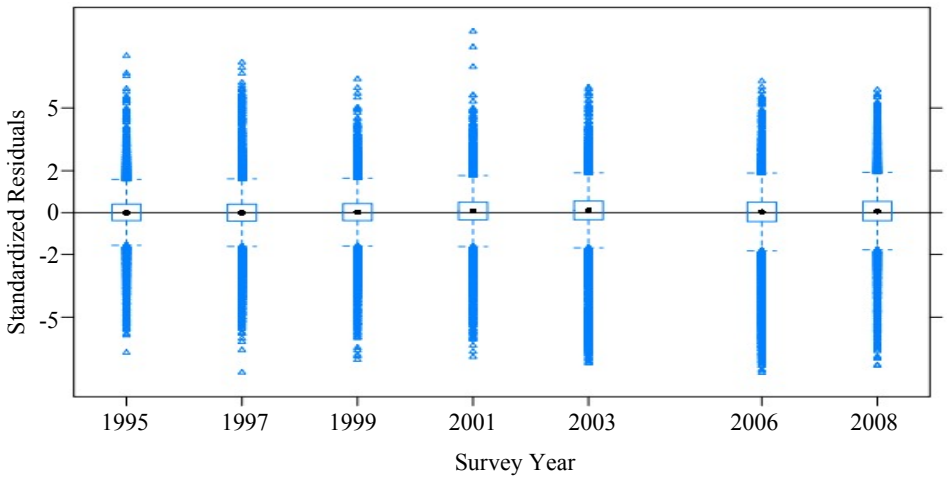


Figure 4.1 Box and Whisker plot of standardized residuals by year

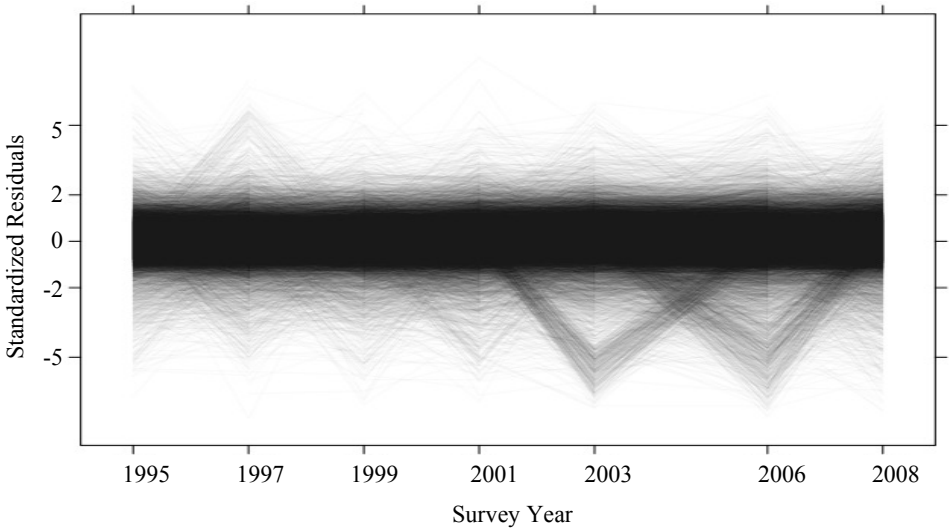


Figure 4.2 Spaghetti plot of standardized residuals

Finally, the plot of fitted values versus observed (which can be found in Carrillo and Karr 2011) also shows a similar story. For most observations the model performs well, apart from those few cases in 2003 and 2006 for whom there is large over-estimation.

5 Conclusions and future research

We have proposed a novel approach to combining different cohorts of a longitudinal survey. The major requirement of our method is that there is a cross-sectional survey weight for each wave, or that one can be built from available information. This weight should allow for statistical inference to the population of interest at the corresponding wave. In that case, our method should perform better than usual estimation procedures (where the auto-correlation is not incorporated) in many practical situations, in particular when there is a high auto-correlation among responses from the same subject.

In general, survey practitioners avoid as much as possible the use of multiple survey weights. However, in the case of rotating panels this is an appealing approach for at least two reasons. On the one hand, it allows for the use of all the available data in a clear and cohesive way in a single analysis procedure. On the other hand, we have shown how readily available cross-sectional survey weights can be directly used for longitudinal analysis, without the need to develop, store, and distribute an additional longitudinal weight or weights.

Our method is directly applicable to any kind of longitudinal survey as long as there are cross-sectional survey weights available (or these can be created) at each wave, and these weights represent the population of interest at the particular wave.

For the theory that we developed about the variance of the estimator proposed, we utilized the (cross-sectional) design weights w_{ij}, which are the inverse of the inclusion probabilities. Yet for the application in our model for salary in the SDR we used the final (cross-sectional) survey weights, which are not the original design weights, but adjusted (in the usual way) weights. This mismatch requires further exploration.

Similarly, in our derivations of the variance, we assumed that the cohorts were independent. However, the SDR does not totally satisfy this assumption for two reasons. Firstly, at any particular wave, the selection of the sample from the old cohorts is not performed independently across cohorts. In order to reduce the number of strata, since 1991 the NSF has collapsed strata

over year of degree receipt for the old cohorts. Additionally, the post-stratification adjustments made to the design weights do not condition over cohort either, and as a result, weights are shared across cohorts. This sampling selection scheme and weighting adjustment procedure violate the independence across cohorts. Some additional calculations (included in the Appendix) have shown that the independence among cohort is not such a crucial requirement for our variance estimation method to produce good approximations, as explained in Section 3.3.1. In future research we plan to evaluate in more detail the impact of this issue.

Acknowledgements

This research was supported by NSF grant SRS-1019244 to the National Institute of Statistical Sciences (NISS). Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation. The authors thank Paul Biemer of RTI International, Stephen Cohen and Nirmala Kannankutty of the National Center for Science and Engineering Statistics at NSF, and Criselda Toto, formerly of NISS, for numerous insightful discussions during the research. We are also grateful to the Associate Editor and two referees for their useful suggestions.

Appendix - Proofs

- To develop an expression for C_ξ, we first simplify $\Psi_s(\beta)\Psi'_U(\beta)$. Let $\boldsymbol{F}_{i(k)} = B_i \mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(k\cdots 3)}$ for $k = 1, 2, 3,$ then we have:

$$N^2\Psi_s(\beta)\Psi'_U(\beta) = \sum_{i\in s} B_i W_i \boldsymbol{e}_i \sum_{i\in U} \boldsymbol{e}'_i \mathrm{I}_i(\mathrm{U}) B'_i = \left[\sum_{i\in s} B_i W_i \boldsymbol{e}_i\right]\left[\sum_{i\in s} \boldsymbol{F}'_{i(1)} + \sum_{i\notin s} \boldsymbol{F}'_{i(1)}\right]$$

$$= \sum_{i\in s} B_i W_i \boldsymbol{e}_i \sum_{i\in s} \boldsymbol{F}'_{i(1)} + \sum_{i\in s} B_i W_i \boldsymbol{e}_i \sum_{i\notin s} \boldsymbol{F}'_{i(1)}$$

$$= \sum_{i\in s} B_i W_i \boldsymbol{e}_i \boldsymbol{e}'_i B'_i + \sum_{i\in s}\sum_{\substack{k\in s \\ k\neq i}} B_i W_i \boldsymbol{e}_i \boldsymbol{e}'_k \mathrm{I}_k(\mathrm{U}) B'_k + \mathrm{A},$$

where $\mathrm{A} = \left(\sum_{i \in s} B_i W_i \boldsymbol{e}_i \right) \left(\sum_{i \notin s} \boldsymbol{F}'_{i(1)} \right)$, and let $\mathrm{B} = \sum_{i \in s} \sum_{\substack{k \in s \\ k \neq i}} B_i W_i \boldsymbol{e}_i \boldsymbol{e}'_k \mathrm{I}_k(\mathrm{U}) B'_k$. The two sums in A are model-independent, $\boldsymbol{e}_i$ and $\boldsymbol{e}'_k$ (in B) are two model-independent terms, and A and B both have model-expectation zero; therefore, $E_\xi[\Psi_s(\boldsymbol{\beta})\, \Psi'_U(\boldsymbol{\beta})] = N^{-2} \sum_{i \in s} B_i W_i E_\xi[\boldsymbol{e}_i \boldsymbol{e}'_i]\, B'_i = N^{-2} \sum_{i \in s} B_i W_i \Sigma_i B'_i = N^{-1} \hat{H}_{\Sigma V}(\boldsymbol{\beta})$; equation (3.9) follows.

- We now develop the expression for $\mathrm{Var}_p[\Psi_s(\boldsymbol{\beta}_N)]$, the design variance of the estimating function; we redefine $B_i = (\partial \boldsymbol{\mu}'_i / \partial \boldsymbol{\beta})\,|_{\boldsymbol{\beta}=\boldsymbol{\beta}_N}\, V_i^{-1}$ and $\boldsymbol{e}_i = \boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}_N)$; then

$$\begin{aligned}
\mathrm{Var}_p[\Psi_s(\boldsymbol{\beta}_N)] &= \mathrm{Var}_p \left(\frac{1}{N} \sum_{i \in s} B_i W_i \boldsymbol{e}_i \right) \\
&= \frac{1}{N^2} \mathrm{Var}_p \left(\sum_{i \in s_{1(1)}} B_i W_i \boldsymbol{e}_i + \sum_{i \in s_{2(2)}} B_i W_i \boldsymbol{e}_i + \sum_{i \in s_{3(3)}} B_i W_i \boldsymbol{e}_i \right) \\
&= \frac{1}{N^2} \mathrm{Var}_p \left(\sum_{i \in s_{1(1)}} B_i W_i \boldsymbol{e}_i \right) + \frac{1}{N^2} \mathrm{Var}_p \left(\sum_{i \in s_{2(2)}} B_i W_i \boldsymbol{e}_i \right) \\
&\quad + \frac{1}{N^2} \mathrm{Var}_p \left(\sum_{i \in s_{3(3)}} B_i W_i \boldsymbol{e}_i \right) = D_{(1)} + D_{(2)} + D_{(3)}, \qquad \text{(A.1)}
\end{aligned}$$

where, for line (A.1), we assume that the (three) cohorts are design-independent. Now, $N^2 D_{(1)} = \mathrm{Var}_p \left[\sum_{i \in U_{1(1)}} B_i W_i\, \mathrm{Diag}\{\boldsymbol{I}_{i(1)}\}\, \boldsymbol{e}_i \right] = \mathrm{Var}_p \left[\sum_{i \in U_{1(1)}} B_i W_i\, \mathrm{Diag}\{\boldsymbol{e}_i\}\, \boldsymbol{I}_{i(1)} \right]$, where $\mathrm{Diag}\{\boldsymbol{e}\}$ is, for a column vector $\boldsymbol{e}$, a diagonal matrix with diagonal entries being the elements of $\boldsymbol{e}$, and $\boldsymbol{I}_{i(1)} = \left(I_i(s_{1(1)}), I_i(s_{2(1)}) I_i(s_{1(1)}), I_i(s_{3(1)}) I_i(s_{2(1)}) I_i(s_{1(1)}) \right)'$. Similarly we can get $N^2 D_{(2)} = \mathrm{Var}_p \left[\sum_{i \in U_{2(2)}} B_i W_i \mathrm{Diag}\{\boldsymbol{e}_i\} \boldsymbol{I}_{i(2)} \right]$ and $N^2 D_{(3)} = \mathrm{Var}_p \left[\sum_{i \in U_{3(3)}} B_i W_i\, \mathrm{Diag}\{\boldsymbol{e}_i\}\, \boldsymbol{I}_{i(3)} \right]$, where $\boldsymbol{I}_{i(2)} = \left(0, I_i(s_{2(2)}), I_i(s_{3(2)}) I_i(s_{2(2)}) \right)'$, and $\boldsymbol{I}_{i(3)} = \left(0, 0, I_i(s_{3(3)}) \right)'$. Now, let us concentrate on $D_{(1)}$; letting $C_i = B_i W_i \mathrm{Diag}\{\boldsymbol{e}_i\}$, we have:

$$
\begin{aligned}
N^2 D_{(1)} &= \mathrm{Var}_p\left(\sum_{i\in U_{1(1)}} C_i \boldsymbol{I}_{i(1)}\right) = \mathrm{Var}\left\{E\left[\sum_{i\in U_{1(1)}} C_i \boldsymbol{I}_{i(1)} \mid s_{1(1)}\right]\right\} \\
&\quad + E\left\{\mathrm{Var}\left[\sum_{i\in U_{1(1)}} C_i \boldsymbol{I}_{i(1)} \mid s_{1(1)}\right]\right\} \\
&= \mathrm{Var}\left\{E\left[E\left(\sum_{i\in U_{1(1)}} C_i \boldsymbol{I}_{i(1)} \mid s_{2(1)}, s_{1(1)}\right) \mid s_{1(1)}\right]\right\} \\
&\quad + E\left\{\mathrm{Var}\left[E\left(\sum_{i\in U_{1(1)}} C_i \boldsymbol{I}_{i(1)} \mid s_{2(1)}, s_{1(1)}\right) \mid s_{1(1)}\right]\right. \\
&\quad \left. + E\left[\mathrm{Var}\left(\sum_{i\in U_{1(1)}} C_i \boldsymbol{I}_{i(1)} \mid s_{2(1)}, s_{1(1)}\right) \mid s_{1(1)}\right]\right\} \\
&= N^2 D_{(1)1} + N^2 D_{(1)2} + N^2 D_{(1)3}. \qquad \text{(A.2)}
\end{aligned}
$$

Let us do each of the terms in (A.2) in turn, beginning with $N^2 D_{1(1)}$, we have:

$$
E\left(\sum_{i\in U_{1(1)}} B_i W_i \mathrm{Diag}\{\boldsymbol{e}_i\} \boldsymbol{I}_{i(1)} \mid s_{2(1)}, s_{1(1)}\right) = \sum_{i\in U_{1(1)}} B_i W_i \mathrm{Diag}\{\boldsymbol{e}_i\} \boldsymbol{I}^{(1)}_{i(1)},
$$

where $\boldsymbol{I}^{(1)}_{i(1)} = (I_i(s_{1(1)}), I_i(s_{2(1)}) I_i(s_{1(1)}), \pi_{i3|s_{2(1)}} I_i(s_{2(1)}) I_i(s_{1(1)}))'$, then

$$
\begin{aligned}
&E\left[E\left(\sum_{i\in U_{1(1)}} C_i \boldsymbol{I}_{i(1)} \mid s_{2(1)}, s_{1(1)}\right) \mid s_{1(1)}\right] \\
&\qquad = \sum_{i\in U_{1(1)}} C_i \boldsymbol{I}^{(2)}_{i(1)} = \sum_{i\in U_{1(1)}} B_i \mathrm{I}^{(1)}_i(\mathrm{U}) \mathrm{Diag}\{\boldsymbol{I}^{(2)}_{i(1)}\} \boldsymbol{e}_i \\
&\qquad = \sum_{i\in U_{1(1)}} F_i \left[\frac{I_i(s_{1(1)})}{\pi_{i1}}, \frac{I_i(s_{1(1)})}{\pi_{i1}}, \frac{I_i(s_{1(1)})}{\pi_{i1}}\right]' \\
&\qquad = \sum_{i\in U_{1(1)}} F_i \mathbf{1}_3 \frac{I_i(s_{1(1)})}{\pi_{i1}} = \sum_{i\in U_{1(1)}} w_{i1(1)} \boldsymbol{F}_{i(1)} I_i(s_{1(1)}),
\end{aligned}
$$

where $\boldsymbol{I}^{(2)}_{i(1)} = (I_i(s_{1(1)}), \pi_{i2|s_{1(1)}} I_i(s_{1(1)}), \pi_{i3|s_{2(1)}} \pi_{i2|s_{1(1)}} I_i(s_{1(1)}))'$, $\mathrm{I}^{(1)}_i(\mathrm{U}) = \mathrm{diag}[I_i(U_1)/\pi_{i1}$, $I_i(U_2) / (\pi_{i1}\pi_{i2|s_{1(1)}})$, $I_i(U_3) / (\pi_{i1}\pi_{i2|s_{1(1)}}\pi_{i3|s_{2(1)}})]$, $F_i = B_i \mathrm{I}_i(\mathrm{U}) \mathrm{Diag}\{\boldsymbol{e}_i\}$, and $\mathbf{1}_3 = (1, 1, 1)'$; this implies that $N^2 D_{(1)1} = \mathrm{Var}\left[\sum_{i\in U_{1(1)}} w_{i1} B_i \mathrm{I}_i(\mathrm{U})\, \boldsymbol{e}_i I_i(s_{1(1)})\right] = \mathrm{Var}\left[\sum_{i\in s_{1(1)}} w_{i1} \boldsymbol{F}_{i(1)}\right]$.

For $N^2 D_{(1)2}$, we have:

$$
\begin{aligned}
&E\left(\sum_{i\in U_{1(1)}} C_i \boldsymbol{I}_{i(1)} \mid s_{2(1)}, s_{1(1)}\right) \\
&\quad = \sum_{i\in U_{1(1)}} B_i W_i \mathrm{Diag}\{\boldsymbol{e}_i\}\boldsymbol{I}_{i(1)}^{(1)} = \sum_{i\in U_{1(1)}} B_i \mathrm{I}_i^{(1)}(\mathrm{U})\mathrm{Diag}\{\boldsymbol{I}_{i(1)}^{(3)}\}\boldsymbol{e}_i \\
&\quad = \sum_{i\in U_{1(1)}} B_i \mathrm{I}_i(\mathrm{U})\mathrm{Diag}\{\boldsymbol{e}_i\}\left[\frac{I_i(s_{1(1)})}{\pi_{i1}}, \frac{I_i(s_{2(1)})I_i(s_{1(1)})}{\pi_{i1}\pi_{i2|s_{1(1)}}}, \frac{I_i(s_{2(1)})I_i(s_{1(1)})}{\pi_{i1}\pi_{i2|s_{1(1)}}}\right]' \\
&\quad = \sum_{i\in s_{1(1)}} w_{i2} B_i \mathrm{I}_i(\mathrm{U})\mathrm{Diag}\{\boldsymbol{e}_i\}[\pi_{i2|s_{1(1)}}, I_i(s_{2(1)}), I_i(s_{2(1)})]',
\end{aligned}
$$

where $\boldsymbol{I}_{i(1)}^{(3)} = (I_i(s_{1(1)}), I_i(s_{2(1)})I_i(s_{1(1)}), \pi_{i3|s_{2(1)}} I_i(s_{2(1)})I_i(s_{1(1)}))'$; then,

$$
\begin{aligned}
&\mathrm{Var}\left[E\left(\sum_{i\in U_{1(1)}} C_i \boldsymbol{I}_{i(1)} \mid s_{2(1)}, s_{1(1)}\right) \mid s_{1(1)}\right] \\
&\quad = \mathrm{Var}\left[\sum_{i\in s_{1(1)}} w_{i2} B_i \mathrm{I}_i(\mathrm{U})\mathrm{Diag}\{\boldsymbol{e}_i\}\boldsymbol{I}_{i(1)}^{(4)} \mid s_{1(1)}\right] \\
&\quad = \mathrm{Var}\left[\sum_{i\in s_{1(1)}} w_{i2} B_i \mathrm{I}_i(\mathrm{U})\mathrm{Diag}\{\boldsymbol{e}_i\}[0, I_i(s_{2(1)}), I_i(s_{2(1)})]' \mid s_{1(1)}\right] \\
&\quad = \mathrm{Var}\left[\sum_{i\in s_{1(1)}} w_{i2} B_i \mathrm{I}_i(\mathrm{U})\mathrm{Diag}\{\boldsymbol{e}_i\} I_i(s_{2(1)})\mathbf{1}_{02} \mid s_{1(1)}\right] \\
&\quad = \mathrm{Var}\left[\sum_{i\in s_{2(1)}} w_{i2} B_i \mathrm{I}_i(\mathrm{U})\boldsymbol{e}_{i(2\cdots 3)} \mid s_{1(1)}\right], \qquad \text{(A.3)}
\end{aligned}
$$

where $\boldsymbol{I}_{i(1)}^{(4)} = [\pi_{i2|s_{1(1)}}, I_i(s_{2(1)}), I_i(s_{2(1)})]'$, $\mathbf{1}_{02} = (0, 1, 1)'$, and line (A.3) is because, conditional on $s_{1(1)}, \pi_{i2|s_{1(1)}}$ is constant and therefore the variance of that component is zero. This means that:

$$
\begin{aligned}
N^2 D_{(1)2} &= E\left\{\mathrm{Var}\left[E\left(\sum_{i\in U_{1(1)}} C_i \boldsymbol{I}_{i(1)} \mid s_{2(1)}, s_{1(1)}\right) \mid s_{1(1)}\right]\right\} \\
&= E\left\{\mathrm{Var}\left[\sum_{i\in s_{2(1)}} w_{i2}\boldsymbol{F}_{i(2)} \mid s_{1(1)}\right]\right\} \\
&= \mathrm{Var}\left[\sum_{i\in s_{2(1)}} w_{i2}\boldsymbol{F}_{i(2)}\right] - \mathrm{Var}\left\{E\left[\sum_{i\in s_{2(1)}} w_{i2}\boldsymbol{F}_{i(2)} \mid s_{1(1)}\right]\right\}
\end{aligned}
$$

$$= \text{Var}\left[\sum_{i\in s_{2(1)}} w_{i2}\boldsymbol{F}_{i(2)}\right] - \text{Var}\left\{E\left[\sum_{i\in s_{2(1)}} w_{i2|s_{1(1)}} w_{i1}\boldsymbol{F}_{i(2)} \mid s_{1(1)}\right]\right\}$$

$$= \text{Var}\left[\sum_{i\in s_{2(1)}} w_{i2}\boldsymbol{F}_{i(2)}\right] - \text{Var}\left\{\sum_{i\in s_{1(1)}} w_{i1}\boldsymbol{F}_{i}\right\}.$$

We can, similarly, show that:

$$N^2 D_{(1)3} = E\left\{E\left[\text{Var}\left(\sum_{i\in U_{1(1)}} B_i W_i \text{Diag}\{\boldsymbol{e}_i\}\boldsymbol{I}_{i(1)} \mid s_{2(1)}, s_{1(1)}\right) \mid s_{1(1)}\right]\right\}$$

$$= E\left\{E\left[\text{Var}\left(\sum_{i\in s_{3(1)}} w_{i3} I_i(s_{2(1)}) I_i(s_{1(1)})\boldsymbol{F}_{i(3)} \mid s_{2(1)}, s_{1(1)}\right) \mid s_{1(1)}\right]\right\}$$

$$= E\left\{\text{Var}\left[\sum_{i\in s_{3(1)}} w_{i3} I_i(s_{2(1)})\boldsymbol{F}_{i(3)} \mid s_{1(1)}\right]\right.$$

$$\left. -\text{Var}\left[E\left(\sum_{i\in s_{3(1)}} w_{i3} I_i(s_{2(1)})\boldsymbol{F}_{i(3)} \mid s_{2(1)}, s_{1(1)}\right) \mid s_{1(1)}\right]\right\}$$

$$= E\left\{\text{Var}\left[\sum_{i\in s_{3(1)}} w_{i3}\boldsymbol{F}_{i(3)} \mid s_{1(1)}\right] - \text{Var}\left[\sum_{i\in s_{2(1)}} w_{i2}\boldsymbol{F}_{i(3)} \mid s_{1(1)}\right]\right\}$$

$$= \text{Var}\left[\sum_{i\in s_{3(1)}} w_{i3}\boldsymbol{F}_{i(3)}\right] - \text{Var}\left[E\left(\sum_{i\in s_{3(1)}} w_{i3}\boldsymbol{F}_{i(3)} \mid s_{1(1)}\right)\right]$$

$$-\text{Var}\left[\sum_{i\in s_{2(1)}} w_{i2}\boldsymbol{F}_{i(3)}\right] + \text{Var}\left[E\left(\sum_{i\in s_{2(1)}} w_{i2}\boldsymbol{F}_{i(3)} \mid s_{1(1)}\right)\right]$$

$$= \text{Var}\left[\sum_{i\in s_{3(1)}} w_{i3}\boldsymbol{F}_{i(3)}\right] - \text{Var}\left[\sum_{i\in s_{1(1)}} w_{i1}\boldsymbol{F}_{i(3)}\right]$$

$$-\text{Var}\left[\sum_{i\in s_{2(1)}} w_{i2}\boldsymbol{F}_{i(3)}\right] + \text{Var}\left[\sum_{i\in s_{1(1)}} w_{i1}\boldsymbol{F}_{i(3)}\right].$$

With similar calculations, we obtain the corresponding expressions for $N^2 D_{(2)}$, $N^2 D_{(2)2}$, $N^2 D_{(2)3}$, and $N^2 D_{(3)} = N^2 D_{(3)3}$.

- Finally, we sketch the development of an expression for $\text{Var}_p[\Psi_s(\boldsymbol{\beta}_N)]$ without assuming independence among cohorts. First, notice that $\Psi_s(\boldsymbol{\beta}_N)$ can be written as:

$$\sum_{i\in s_1} B_i \mathrm{I}_i(\mathrm{U}) \begin{bmatrix} w_{i1} & & \mathrm{O} \\ & w_{i2} & \\ \mathrm{O} & & w_{i3} \end{bmatrix} \begin{bmatrix} e_{i1} \\ e_{i2} \\ e_{i3} \end{bmatrix} + \sum_{i\in s_2} B_i \mathrm{I}_i(\mathrm{U}) \begin{bmatrix} 0 & & \mathrm{O} \\ & w_{i2} & \\ \mathrm{O} & & w_{i3} \end{bmatrix} \begin{bmatrix} 0 \\ e_{i2} \\ e_{i3} \end{bmatrix}$$

$$- \sum_{i\in s_1} B_i \mathrm{I}_i(\mathrm{U}) \begin{bmatrix} 0 & & \mathrm{O} \\ & w_{i2} & \\ \mathrm{O} & & w_{i3} \end{bmatrix} \begin{bmatrix} 0 \\ e_{i2} \\ e_{i3} \end{bmatrix}$$

$$+ \sum_{i\in s_3} B_i \mathrm{I}_i(\mathrm{U}) \begin{bmatrix} 0 & & \mathrm{O} \\ & 0 & \\ \mathrm{O} & & w_{i3} \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ e_{i3} \end{bmatrix} - \sum_{i\in s_2} B_i \mathrm{I}_i(\mathrm{U}) \begin{bmatrix} 0 & & \mathrm{O} \\ & 0 & \\ \mathrm{O} & & w_{i3} \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ e_{i3} \end{bmatrix}$$

$$= \sum_{i\in s_1} w_{i1} B_i \mathrm{I}_i(\mathrm{U}) \boldsymbol{e}_i - \sum_{i\in s_1} w_{i1} B_i \mathrm{I}_i(\mathrm{U}) \begin{bmatrix} 0 \\ e_{i2} \\ e_{i3} \end{bmatrix}$$

$$+ \sum_{i\in s_2} w_{i2} B_i \mathrm{I}_i(\mathrm{U}) \begin{bmatrix} 0 \\ e_{i2} \\ e_{i3} \end{bmatrix} - \sum_{i\in s_2} w_{i2} B_i \mathrm{I}_i(\mathrm{U}) \begin{bmatrix} 0 \\ 0 \\ e_{i3} \end{bmatrix}$$

$$+ \sum_{i\in s_3} w_{i3} B_i \mathrm{I}_i(\mathrm{U}) \begin{bmatrix} 0 \\ 0 \\ e_{i3} \end{bmatrix};$$

letting $\boldsymbol{z}_i = B_i \mathrm{I}_i(\mathrm{U}) \boldsymbol{e}_i$, $\boldsymbol{z}_{i(2\cdots3)} = B_i \mathrm{I}_i(\mathrm{U})[0, e_{i2}, e_{i3}]'$, and $\boldsymbol{z}_{i(3\cdots3)} = B_i \mathrm{I}_i(\mathrm{U})[0, 0, e_{i3}]'$, $\mathrm{Var}_p[\Psi_s(\boldsymbol{\beta}_N)]$ can be expanded as:

$$\mathrm{Var}_p \left[\sum_{i\in s_1} w_{i1} \boldsymbol{z}_i \right] + \mathrm{Var}_p \left[\sum_{i\in s_1} w_{i1} \boldsymbol{z}_{i(2\cdots3)} \right]$$
$$+ \mathrm{Var}_p \left[\sum_{i\in s_2} w_{i2} \boldsymbol{z}_{i(2\cdots3)} \right] + \mathrm{Var}_p \left[\sum_{i\in s_2} w_{i2} \boldsymbol{z}_{i(3\cdots3)} \right]$$
$$+ \mathrm{Var}_p \left[\sum_{i\in s_3} w_{i3} \boldsymbol{z}_{i(3\cdots3)} \right] - 2\mathrm{Cov}_p \left[\sum_{i\in s_1} w_{i1} \boldsymbol{z}_i, \sum_{i\in s_1} w_{i1} \boldsymbol{z}_{i(2\cdots3)} \right]$$
$$+ 2\mathrm{Cov}_p \left[\sum_{i\in s_1} w_{i1} \boldsymbol{z}_i, \sum_{i\in s_2} w_{i2} \boldsymbol{z}_{i(2\cdots3)} \right]$$
$$- 2\mathrm{Cov}_p \left[\sum_{i\in s_1} w_{i1} \boldsymbol{z}_i, \sum_{i\in s_2} w_{i2} \boldsymbol{z}_{i(3\cdots3)} \right] + 2\mathrm{Cov}_p \left[\sum_{i\in s_1} w_{i1} \boldsymbol{z}_i, \sum_{i\in s_3} w_{i3} \boldsymbol{z}_{i(3\cdots3)} \right]$$
$$- 2\mathrm{Cov}_p \left[\sum_{i\in s_1} w_{i1} \boldsymbol{z}_{i(2\cdots3)}, \sum_{i\in s_2} w_{i2} \boldsymbol{z}_{i(2\cdots3)} \right] + 2\mathrm{Cov}_p \left[\sum_{i\in s_1} w_{i1} \boldsymbol{z}_{i(2\cdots3)}, \sum_{i\in s_2} w_{i2} \boldsymbol{z}_{i(3\cdots3)} \right]$$
$$- 2\mathrm{Cov}_p \left[\sum_{i\in s_1} w_{i1} \boldsymbol{z}_{i(2\cdots3)}, \sum_{i\in s_3} w_{i3} \boldsymbol{z}_{i(3\cdots3)} \right] - 2\mathrm{Cov}_p \left[\sum_{i\in s_2} w_{i2} \boldsymbol{z}_{i(2\cdots3)}, \sum_{i\in s_2} w_{i2} \boldsymbol{z}_{i(3\cdots3)} \right]$$
$$+ 2\mathrm{Cov}_p \left[\sum_{i\in s_2} w_{i2} \boldsymbol{z}_{i(2\cdots3)}, \sum_{i\in s_3} w_{i3} \boldsymbol{z}_{i(3\cdots3)} \right] - 2\mathrm{Cov}_p \left[\sum_{i\in s_2} w_{i2} \boldsymbol{z}_{i(3\cdots3)}, \sum_{i\in s_3} w_{i3} \boldsymbol{z}_{i(3\cdots3)} \right]$$

$$
\begin{aligned}
&= \mathrm{Var}_p\left[\sum_{i\in s_1} w_{i1}\mathbf{z}_i\right] + \mathrm{Var}_p\left[\sum_{i\in s_2} w_{i2}\mathbf{z}_{i(2\cdots 3)}\right] - \mathrm{Var}_p\left[\sum_{i\in s_1} w_{i1}\mathbf{z}_{i(2\cdots 3)}\right] \\
&+ \mathrm{Var}_p\left[\sum_{i\in s_3} w_{i3}\mathbf{z}_{i(3\cdots 3)}\right] - \mathrm{Var}_p\left[\sum_{i\in s_2} w_{i2}\mathbf{z}_{i(3\cdots 3)}\right] \\
&+ 2\mathrm{Cov}_p\left[\sum_{i\in s_1} w_{i1}\mathbf{z}_{i(1\cdots 1)}, \sum_{i\in s_2} w_{i2}\mathbf{z}_{i(2\cdots 3)}\right] - 2\mathrm{Cov}_p\left[\sum_{i\in s_1} w_{i1}\mathbf{z}_{i(1\cdots 1)}, \sum_{i\in s_1} w_{i1}\mathbf{z}_{i(2\cdots 3)}\right] \\
&+ 2\mathrm{Cov}_p\left[\sum_{i\in s_1} w_{i1}\mathbf{z}_{i(1\cdots 2)}, \sum_{i\in s_3} w_{i3}\mathbf{z}_{i(3\cdots 3)}\right] - 2\mathrm{Cov}_p\left[\sum_{i\in s_1} w_{i1}\mathbf{z}_{i(1\cdots 2)}, \sum_{i\in s_2} w_{i2}\mathbf{z}_{i(3\cdots 3)}\right] \\
&+ 2\mathrm{Cov}_p\left[\sum_{i\in s_1} w_{i1}\mathbf{z}_{i(2\cdots 2)}, \sum_{i\in s_2} w_{i2}\mathbf{z}_{i(3\cdots 3)}\right] - 2\mathrm{Cov}_p\left[\sum_{i\in s_1} w_{i1}\mathbf{z}_{i(2\cdots 2)}, \sum_{i\in s_3} w_{i3}\mathbf{z}_{i(3\cdots 3)}\right] \\
&+ 2\mathrm{Cov}_p\left[\sum_{i\in s_2} w_{i2}\mathbf{z}_{i(2\cdots 2)}, \sum_{i\in s_3} w_{i3}\mathbf{z}_{i(3\cdots 3)}\right] - 2\mathrm{Cov}_p\left[\sum_{i\in s_2} w_{i2}\mathbf{z}_{i(2\cdots 2)}, \sum_{i\in s_2} w_{i2}\mathbf{z}_{i(3\cdots 3)}\right].
\end{aligned}
\tag{A.4}
$$

In this last expression, the first thing we notice is that *all* the diagonal elements in *all* the covariance terms are exactly equal to zero; this means that whether or not the cohorts are independent of one another, expression (3.13) is exact for the variance terms.

To analyze the importance of the covariance terms, we concentrate on the term in line (A.4); the conclusion for the other terms is the same; note that this term can be written as:

$$
2\mathrm{Cov}_p\left[\sum_{i\in s_1} w_{i1}\frac{\partial\boldsymbol{\mu}_i'}{\partial\boldsymbol{\beta}}V_i^{-1}\begin{pmatrix} e_{i1}\\ e_{i2}\\ 0\end{pmatrix}, \sum_{i\in s_3} w_{i3}\frac{\partial\boldsymbol{\mu}_i'}{\partial\boldsymbol{\beta}}V_i^{-1}\begin{pmatrix} 0\\ 0\\ e_{i3}\end{pmatrix} - \sum_{i\in s_2} w_{i2}\frac{\partial\boldsymbol{\mu}_i'}{\partial\boldsymbol{\beta}}V_i^{-1}\begin{pmatrix} 0\\ 0\\ e_{i3}\end{pmatrix}\right];
$$

Property 3.1 states that if the *cohorts* are design-independent, all the covariance terms are exactly equal to zero. In addition to that, from this last expression we conclude, trivially, that if the *waves* are design-independent, all the covariance terms are equal to zero too. This formula for the term in line (A.4) also implies that if the individual weights do not vary greatly between consecutive waves, and there is a high overlap between consecutive waves, the covariance terms are not too large. Finally, if the overlap is small, it is reasonable to assume design-independence between the waves, and then the covariance terms can be safely approximated by zero.

References

Ardilly, P., and Lavallée, P. (2007). Weighting in rotating samples: The SILC survey in France. *Survey Methodology*, 33, 2, 131-137.

Berger, Y.G. (2004a). Variance estimation for change: An evaluation based upon the 2000 finnish labour force survey. Proceedings. European Conference on Quality and Methodology in Official Statistics.

Berger, Y.G. (2004b). Variance estimation for measures of change in probability sampling. *The Canadian Journal of Statistics*, 32, 4, 451-467.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

Carrillo, I.A., Chen, J. and Wu, C. (2010). The pseudo-GEE approach to the analysis of longitudinal surveys. *The Canadian Journal of Statistics*, 38, 4, 540-554.

Carrillo, I.A., Chen, J. and Wu, C. (2011). A pseudo-GEE approach to analyzing longitudinal surveys under imputation for missing responses. *Journal of Official Statistics*, 27, 2, 255-277.

Carrillo, I.A., and Karr, A.F. (2011). Combining cohorts in longitudinal surveys. Technical Report 180, National Institute of Statistical Sciences, Research Triangle Park, NC. URL http://www.niss.org/sites/default/files/tr180.pdf.

Carrillo, I.A., and Karr, A.F. (2012). Estimating change with multi-cohort longitudinal surveys. In preparation.

Cox, B.G., Grigorian, K., Wang, R. and Harter, R. (2010). 2008 Survey of Doctorate Recipients Weighting Implementation Report, document prepared by the National Opinion Research Center (NORC) for the National Science Foundation (NSF).

Diggle, P., Heagerty, P., Liang, K.-Y. and Zeger, S. (2002). *Analysis of Longitudinal Data*, 2nd Edition. Oxford University Press, New York.

Hedeker, D., and Gibbons, R.D. (2006). *Longitudinal Data Analysis*. Wiley Series in Probability and Statistics. New Jersey: John Wiley & Sons, Inc., Hoboken.

Hirano, K., Imbens, G.W., Ridder, G. and Rubin, D.B. (2001). Combining panel data sets with attrition and refreshment samples. *Econometrica*, 69, 6, 1645-1659.

Hu, F., and Kalbfleisch, J.D. (2000). The estimating function bootstrap (Pkg: P449-495). *The Canadian Journal of Statistics*, 28, 3, 449-481.

Larsen, M.D., Qing, S., Zhou, B. and Foulkes, M.A. (2011). Calibration estimation and longitudinal survey weights: Application to the NSF Survey of Doctorate Recipients. In *Proceedings of the Survey Research Method Section*, American Statistical Association, 1360-1374.

Liang, K.-Y., and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.

Lohr, S. (2007). Recent developments in multiple frame surveys. In *Joint Statistical Meeting of the American Statistical Association*, 3257-3264.

McLaren, C.H., and Steel, D.G. (2000). The impact of different rotation patterns on the sampling variance of seasonally adjusted and trend estimates. *Survey Methodology*, 26, 2, 163-172.

National Science Foundation, National Center for Science and Engineering Statistics (2012). Survey of doctorate recipients. http://www.nsf.gov/statistics/srvydoctoratework/, accessed Feb. 09 2012.

Nevo, A. (2003). Using weights to adjust for sample selection when auxiliary information is available. *Journal of Business & Economic Statistics*, 21, 1, 43-52.

Qualité, L., and Tillé, Y. (2008). Variance estimation of changes in repeated surveys and its application to the Swiss survey of value added. *Survey Methodology*, 34, 2, 173-181.

Rao, J.N.K., and Wu, C. (2010). Pseudo-empirical likelihood inference for multiple frame surveys. *Journal of the American Statistical Association*, 105, 492, 1494-1503.

Roberts, G., Binder, D., Kovačević, M., Pantel, M. and Phillips, O. (2003). Using an estimating function bootstrap approach for obtaining variance estimates when modelling complex health survey data. *Proceedings of the Survey Methods Section, Statistical Society of Canada*, Halifax.

Robins, J.M., Rotnitzky, A. and Zhao, L.P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90, 106-121.

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

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Smith, P., Lynn, P. and Elliot, D. (2009). Sample design for longitudinal surveys. In *Methodology of Longitudinal Surveys*, (Ed., P. Lynn). Wiley, Chichester, Chapter 2, 21-33.

Song, P.X.-K. (2007). *Correlated Data Analysis: Modeling, Analytics, and Applications*. Springer Series in Statistics. New York: Springer.

Steel, D., and McLaren, C. (2007). Design and analysis of repeated surveys. Keynote lecture. International Conference on Quality Management of Official Statistics, Korea.

Vieira, M.D.T. (2009). Analysis of Longitudinal Survey Data: Allowing for the Complex Survey Design in Covariance Structure Models. VDM Verlag.

Vieira, M.D.T., and Skinner, C.J. (2008). Estimating models for panel survey data under complex sampling. *Journal of Official Statistics*, 24, 3, 343-364.

Survey Methodology, June 2013
Vol. 39, No. 1, pp. 183-215
Statistics Canada, Catalogue No. 12-001-X

Indirect sampling applied to skewed populations

Pierre Lavallée and Sébastien Labelle-Blanchet[1]

Abstract

Indirect Sampling is used when the sampling frame is not the same as the target population, but related to the latter. The estimation process for Indirect Sampling is carried out using the Generalised Weight Share Method (GWSM), which is an unbiased procedure (see Lavallée 2002, 2007). For business surveys, Indirect Sampling is applied as follows: the sampling frame is one of establishments, while the target population is one of enterprises. Enterprises are selected through their establishments. This allows stratifying according to the establishment characteristics, rather than those associated with enterprises. Because the variables of interest of establishments are generally highly skewed (a small portion of the establishments covers the major portion of the economy), the GWSM results in unbiased estimates, but their variance can be large. The purpose of this paper is to suggest some adjustments to the weights to reduce the variance of the estimates in the context of skewed populations, while keeping the method unbiased. After a brief overview of Indirect Sampling and the GWSM, we describe the required adjustments to the GWSM. The estimates produced with these adjustments are compared to those from the original GWSM, via a small numerical example, and using real data originating from the Statistics Canada's Business Register.

Key Words: Generalised weight share method; Weighted links; Weak optimality.

1 Introduction

Business surveys differ in a number of ways from social surveys. One is that data associated with a business frame are highly skewed, whereas those associated with social surveys are much more homogenous. The sampling of businesses takes place by transforming operating structures into standardized units known as statistical units. These are represented as a hierarchy, or series of levels that allow subsequent integration of the various data items available at different levels within the organization. The number of levels within the hierarchy differs between statistical agencies. For example, the Canadian Business Register has four levels: enterprise, company, establishment and location. The statistical enterprise corresponds to the legal unit in most cases. The statistical establishment is, in most cases, equivalent to a profit centre and provides data on

1. Pierre Lavallée, Business Survey Methods Division, Statistics Canada. E-mail: pierre.lavallee@statcan.gc.ca; Sébastien Labelle-Blanchet, Household Survey Methods Division, Statistics Canada. E-mail: sebastien.labelle-blanchet@statcan.gc.ca.

the value of output, the cost of inputs and labour. These data are sufficient to compute value added (profit, salary and wages). In this paper, we will consider only two levels of the hierarchy: enterprise and establishment, as defined by the Canadian Business Register. For more information, one can consult Statistics Canada (2010).

For selecting the sample, stratification is often performed at the establishment level. This allows to control for geographical (*e.g.*, by stratifying by province), industrial (*e.g.*, by stratifying by the industrial activity), and size representativeness (*e.g.*, by stratifying by revenue classes or by the number of employees). Controlling for such representativeness is not possible if stratification is performed at the enterprise level. However, in addition to establishment-based statistics, statistics at the enterprise level are often required. Therefore, to achieve these two goals, we perform the sample selection at the establishment level, and once the sample of establishments is obtained, we extend the sample to the complete set of establishments belonging to the enterprises of the initially selected establishments. Note that selecting enterprises through the selection of establishments is like selecting clusters through their components. This procedure allows producing estimates at the establishment and the enterprise levels, as well as some reduction in collection costs by selecting clusters of establishments.

One way to consider the production of enterprise-level estimates using a sample of establishments is by viewing the sampling frame and the target population separately. The former is a set of establishments, while the latter is a set of enterprises corresponding to clusters of establishments. When the sampling frame is not the same as the target population, but still related to the latter, we are in a situation of *Indirect Sampling* (see Lavallée 2002 (French version), 2007 (English version)). More formally, we wish to produce an estimate for a target population U^B, using a sampling frame U^A, which is somehow linked to U^B. We select a sample s^A from U^A to produce an estimate for U^B using the existing links between the two populations. To produce unbiased estimates of quantities of interest (*e.g.*, totals or means) for the target population U^B using s^A, we obtain estimation weights using the *Generalised Weight Share Method* (GWSM).

Although the theory of the GWSM is widely developed (see Lavallée 2002, 2007), its application for business surveys presents some difficulties. Indeed, while yielding unbiased estimates, the GWSM tends to lead to large variances. This lack of precision is due to the skewness of the population, *i.e.*, a small number of establishments cover a major portion of the economy.

The purpose of this paper is to suggest some adjustments to the estimation weights to reduce the variance of the estimates in the context of skewed populations, while keeping the method unbiased. After a brief overview of Indirect Sampling and the GWSM in Section 2, we present the problem with skewed population in Section 3. We describe the proposed adjustments to be done to the GWSM in Section 4. In Section 5, the estimates produced with these adjustments are compared to those from the original GWSM using a small numerical example, and using real data that come from the Statistics Canada's Business Register. A brief conclusion is presented in Section 6.

2 Indirect sampling and the GWSM

In this section, we provide an overview of Indirect Sampling and the GWSM. Although Indirect Sampling has been developed for any type of sample design, we will focus on stratified Simple Random Sampling Without Replacement (SRSWoR), since this sampling design is the most commonly used for business surveys.

Let the population U^A of M^A establishments be stratified in H strata, where stratum h contains M_h^A establishments. In each stratum h, we select a sample s_h^A of m_h^A establishments using SRSWoR. Let $s^A = \bigcup_{h=1}^{H} s_h^A$ and $m^A = \sum_{h=1}^{H} m_h^A$. The target population U^B contains N^B enterprises, where enterprise i contains those M_i^B establishments of U^A. This population can also be viewed as a population of M^B establishments, where each establishment k belongs to an enterprise i, with $M^B = \sum_{i=1}^{N^B} M_i^B$.

We wish to produce an estimate for the target population U^B, using the sampling frame U^A along with the existing links between the two populations. The links between population U^A and population U^B are identified by the indicator variable $l_{j,i}$, where $l_{j,i} = 1$ if there exists a link between establishment $j \in U^A$ and enterprise $i \in U^B$, and 0 otherwise. In the present case, $l_{j,i} = 1$ if the establishment j of U^A belongs to enterprise i of U^B, and 0 otherwise. Because each establishment can belong to only one enterprise, the links between U^A and U^B are many-to-one or one-to-one. Therefore, we have $L_j^A = \sum_{i=1}^{N^B} l_{j,i} = 1$, $L_i^B = \sum_{j=1}^{M^A} l_{j,i} = M_i^B$, for all establishments $j \in U^A$ and for all enterprise $i \in U^B$.

Steps for Indirect Sampling:

1. For each establishment j selected in s^A, we identify the corresponding enterprise i of U^B.
2. For each enterprise i identified, we assume that we can set up the list U_i^B of all M_i^B establishments of this enterprise.
3. For each enterprise i identified, we survey *all* M_i^B establishments of the enterprise.
4. At the end, we obtain a sample s^B of n^B enterprises, and this sample contains $m^B = \sum_{i=1}^{n^B} M_i^B$ establishments.

For all the establishments k linked to enterprises $i \in s^B$, we measure a variable of interest y_{ik}. We want to estimate the total $Y = \sum_{i=1}^{N^B} \sum_{k=1}^{M_i^B} y_{ik} = \sum_{i=1}^{N^B} Y_i$ for the target population U^B. Note that the collection process of Indirect Sampling results in a number of surveyed establishments that is much larger than the number of establishments in the initial sample s^A. We initially sample m^A establishments in s^A, and end up with sampling $m^B = \sum_{i=1}^{n^B} M_i^B$ establishments, where $m^B \geq m^A$.

In practice, it can happen that some enterprises only provide their data at the enterprise level. That is, we obtain the values $Y_i = \sum_{k=1}^{M_i^B} y_{ik}$ for $i \in s^B$, but not the values y_{ik} measured at the establishment level. As we will see, this does not create problems for global estimates, but it might create difficulties for some detailed estimates. When this occurs, a disaggregation (also

called allocation) of the enterprise values to the establishment level is performed mainly based on subject matter expertise (see for example, Delorme 2000).

With indirect sampling, nonresponse can be present within the sample s^A selected from U^A, or within the units (enterprises or establishments) identified to be surveyed within U^B. Since the units in population U^B are in fact surveyed by cluster (recall that enterprises are clusters of establishments), there are two types of nonresponse from U^B: cluster nonresponse and unit nonresponse. Cluster nonresponse refers to the case where the variable of interest y is not measured for any of the establishments of the enterprises selected in the survey. Unit nonresponse occurs when one or more establishments of the enterprise, but not all, did not respond. With Indirect Sampling, there is also another form of nonresponse that comes from the problem of identifying some of the links. This type of nonresponse is associated with the situation where it is impossible to determine whether an establishment k of an enterprise i of U^B is linked or not to an establishment j of U^A. This is referred to as the problem of links identification. Lavallée (2002, 2007) proposed solutions to correct these types of nonresponse based on weight adjustments. To restrict the scope of the present paper, we will assume that nonresponse does not occur at any level.

According to the GWSM, to estimate the total Y, we use the estimator

$$\hat{Y} = \sum_{i=1}^{n^B} w_i Y_i \tag{2.1}$$

where n^B is the number of surveyed enterprises. The weights obtained from the GWSM are given by

$$w_i = \sum_{j=1}^{M^A} \frac{t_j^A}{\pi_j^A} \frac{l_{j,i}}{L_i^B} \tag{2.2}$$

where $t_j^A = 1$ if $j \in s^A$, 0 otherwise, and π_j^A is the selection probability of establishment j. In the present case, we have $\pi_j^A = m_h^A / M_h^A$ for $j \in h$. It should be noted that the weights (2.2) do

not correspond, in general, to the selection probabilities π_i^B of the enterprises i. Using (2.2), we can rewrite estimator (2.1) as

$$\hat{Y} = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} Z_j \tag{2.3}$$

where

$$Z_j = \sum_{i=1}^{N^B} \frac{Y_i}{L_i^B} l_{j,i}. \tag{2.4}$$

Because of the many-to-one correspondence between U^A and U^B, we have

$$w_i = \frac{1}{M_i^B} \sum_{j=1}^{M_i^B} \frac{t_j^A}{\pi_j^A}. \tag{2.5}$$

In addition, the variable Z_j of (2.4) can be written as $Z_j = Y_i / M_i^B = \bar{Y}_i$, for $j \in i$, which is the average of the M_i^B establishments belonging to enterprise i. We thus have

$$\hat{Y} = \sum_{h=1}^{H} \frac{M_h^A}{m_h^A} \sum_{j=1}^{m_h^A} Z_{hj} \tag{2.6}$$

where $Z_{hj} = Y_i / M_i^B = \bar{Y}_i$, for $j \in i$.

One can prove that estimator (2.1) (and therefore (2.3) and (2.6)) is unbiased for Y (see Lavallée 2002, 2007). Note that estimator $\hat{Y}$ is in fact only a Horvitz-Thompson estimator where the variable of interest is the variable Z_{hj}. In the case of stratified SRSWoR, its variance is given by

$$\mathrm{Var}(\hat{Y}) = \sum_{h=1}^{H} M_h^A \left(\frac{M_h^A - m_h^A}{m_h^A} \right) S_{Z,h}^2 \tag{2.7}$$

where $S_{Z,h}^2 = \sum_{j=1}^{M_h^A} \left(Z_{hj} - \bar{Z}_h \right)^2 \Big/ (M_h^A - 1)$ and $\bar{Z}_h = \sum_{j=1}^{M_h^A} Z_{hj} \Big/ M_h^A$. The variance $\mathrm{Var}(\hat{Y})$ can be estimated using the classical estimator for stratified SRSWoR, or by other variance estimators

proposed in the scientific literature, such as Jackknife and Bootstrap estimators. See Wolter (2007) or Särndal, Swensson and Wretman (1992).

The precision of the estimates produced using the GWSM depends solely on the variance because the estimator (2.1) (and therefore (2.3) and (2.6)) is unbiased. Looking at equation (2.7), we find that the precision depends, as in the classical case, on the sample sizes and sampling fractions used to select s^A, but also on the variability of the derived variables Z. Since $Z_{hj} = Y_i / M_i^B = \bar{Y}_i$, for $j \in i$, the value of Z_{hj} is the same for all establishments j of a given enterprise i. That is, the enterprise total Y_i is shared equally among its establishments. If all the establishments of an enterprise belong to the same stratum, the variability of the variables Z within a stratum will only depend on the difference between the average values of a limited number of enterprises, which might make the variability to be relatively small. On the other hand, if the establishments of an enterprise belong to different strata, the variability of the variables Z within a stratum will depend on the difference between up to as many enterprises as there are establishments, which might result in a quite large variability. Because of the skewness of the population of establishments and the stratification applied to U^A, the latter case is the one that is most likely to occur.

It is interesting to see that the present version of Indirect Sampling (together with the GWSM) corresponds mathematically to *Adaptive Cluster Sampling* presented by Thompson (1990, 1991, 1992, 2002) and Thompson and Seber (1996). With Adaptive Cluster Sampling, a sample of establishments would first be selected, and a collection strategy would then be performed to survey all establishments of the enterprises identified by the initial sample of selected establishments. Typically, the collection strategy would be to expand the sample of establishments by visiting them sequentially, until all establishments of the same enterprises are covered. With Indirect Sampling, the collection strategy is not specified, but at the end of the collection process, the complete set of establishments of the selected enterprises is assumed to be surveyed. The estimator related to Adaptive Cluster Sampling can be proved the same as

estimator (2.1) obtained through the GWSM (see Lavallée 2002, 2007). Note that the two sampling designs happen to be mathematically equivalent only in some particular cases. This is the case in the present paper when estimator (2.1) is used. When the weighted links (see next section) are used, the GWSM turns out to produce a different estimator than the one related to Adaptive Cluster Sampling. As well, when the links between populations U^A and U^B are many-to-many, Indirect Sampling and Adaptive Cluster Sampling are no longer equivalent.

2.1 Use of weighted links

The indicator variable $l_{j,i}$ simply indicates whether there is a link between establishments j and enterprise i from populations U^A and U^B, respectively. It is however possible to replace the indicator variable $l_{j,i}$ with any quantitative variable $\theta_{j,i}$ representing the importance that we want to give to the link $l_{j,i}$. That is, there is no problem with generalising the indicator variable l defined on $\{0,1\}$ with a quantitative variable θ defined on $[0,+\infty[$, the set of non-negative real numbers. In this case, a value of $\theta_{j,i} = 0$ amounts to a link $l_{j,i} = 0$. The theory developed around the GWSM remains valid. For instance, the resulting estimator is still unbiased. As it will be seen later, choosing appropriate values for the weighted links $\theta_{j,i}$ will be the basis for methods that aim to reduce the variance of the estimates obtained through the GWSM.

Let $\tilde{\theta}_{j,i} = \theta_{j,i} / \theta_i^B$ where $\theta_i^B = \sum_{j=1}^{M^A} \theta_{j,i}$. From (2.2), we define

$$w_i^{\theta} = \sum_{j=1}^{M_i^B} \frac{t_j^A}{\pi_j^A} \tilde{\theta}_{j,i}. \tag{2.8}$$

Using (2.8), we can modify estimator (2.6) as

$$\hat{Y}_{\theta} = \sum_{h=1}^{H} \frac{M_h^A}{m_h^A} \sum_{j=1}^{m_h^A} Z_{hj}^{\theta} \tag{2.9}$$

where

$$Z_{hj}^{\theta} = \sum_{i=1}^{N^B} \tilde{\theta}_{j,i} Y_i \tag{2.10}$$

for $j \in h$. Because of the many-to-one correspondence between U^A and U^B, the variable Z_{hj}^{θ} in (2.10) is a weighted portion of the total Y_i of the M_i^B establishments belonging to enterprise i. The variance of (2.9) is obtained by replacing Z_j by Z_j^{θ} in (2.7):

$$\mathrm{Var}(\hat{Y}_{\theta}) = \sum_{h=1}^{H} M_h^A \left(\frac{M_h^A - m_h^A}{m_h^A} \right) S_{\theta Zh}^2 \tag{2.11}$$

where $S_{\theta Zh}^2 = \sum_{j=1}^{M_h^A} \left(Z_{hj}^{\theta} - \bar{Z}_h^{\theta} \right)^2 \Big/ (M_h^A - 1)$ and $\bar{Z}_h^{\theta} = \sum_{j=1}^{M_h^A} Z_{hj}^{\theta} \Big/ M_h^A$.

2.2 Using optimal weighted links

The GWSM offers a simple solution for obtaining an estimation weight w_i for each surveyed enterprise i. However, the resulting estimator $\hat{Y}$ given by (2.1) and (2.3) resulting from the default use of the GWSM is not always the one that has the smallest variance. It is possible to improve it by determining optimal weights for the links $\theta_{j,i}$. This problem has been solved by Deville and Lavallée (2006).

We pointed out earlier that the variance (2.7) depends on the variability of the derived variables Z. Without weighted links, *i.e.*, with $Z_{hj} = Y_i / M_i^B = \bar{Y}_i$, for $j \in i$, the value of Z_{hj} is the same for all establishments j of a given enterprise i. Because it is likely that the establishments of an enterprise belong to different strata, the variability of the variables Z within a stratum will depend on the difference between up to as many enterprises as there are establishments. Moreover, a given enterprise i will provide the same value of Z to all its establishments j since $Z_{hj} = \bar{Y}_i$. Therefore, whether an establishment is part of a stratum of "large" or "small" units (with respect to some size measure) or not, this establishment will receive the average value of its owning enterprise. This will contribute to increase the variability within strata, and thus, to increase the variance (2.7). The idea behind the use of weighted links is to share the value of the enterprise total Y_i unequally between its establishments. Searching for optimal weighted links is to seek for sharing the value of the enterprise total Y_i in such a way that the variance (2.11) will be minimal.

Deville and Lavallée (2006) obtained an estimator that has a variance less than or equal to that of the original estimator $\hat{Y}$. As mentioned earlier, estimator $\hat{Y}_\theta$ given by (2.9) will still provide unbiased estimates. Now, the variance (2.11) of this estimator depends on the weighted links $\theta_{j,i}$. The problem is then to find at least one set of values $\theta_{j,i}$ such that the variance of the estimator $\hat{Y}_\theta$ is minimal. That is, for the $\theta_{j,i}$ that are greater than 0, we want to determine the values such that we obtain the most precise estimator $\hat{Y}_\theta$. The solution to this problem is obtained by minimising the variance (2.11) with respect to the weighted links $\theta_{j,i}$, which is a relatively standard and simple problem to solve. However, the solution is not trivial to write, and it often depends on the variable of interest y.

If the optimal weighted links $\theta_{j,i}^{\text{opt}}$ depend on the variable of interest y, then the weights w_i^θ will also depend on y. This means that a different set of weights will need to be computed for each variable of interest. To overcome this problem, Deville and Lavallée (2006) defined *weak optimality*, which corresponds to minimising the variance (2.11) for a very specific choice of a variable of interest: $Y_i = 1$ for an enterprise i of U^B and $Y_{i'} = 0$ for all other enterprises i' of $U^B (i' \neq i)$. The resulting weak-optimal weighted links do not involve, per se, the variable y and they turn out to be relatively easy to compute, *i.e.*, they can be obtained as a closed-form solution, without the need of numerical computations. In addition, if some conditions given by Deville and Lavallée (2006) are satisfied, then weak-optimality corresponds to *strong optimality independent of* y. That is, the weighted links $\theta_{j,i}^{w-\text{opt}}$ obtained through weak optimality correspond to the optimal weighted links $\theta_{j,i}^{\text{opt}}$ obtained by minimising (2.11), and they do not depend on the variable of interest y. Unfortunately, these conditions are rarely satisfied in practice, even for simple sampling designs such as SRSWoR.

Assuming SRSWoR *without* stratification, it can be shown that the weak-optimal weighted links are given by $\tilde{\theta}_{j,i}^{w-\text{opt}} = \theta_{j,i}^{w-\text{opt}} \Big/ \sum_{j=1}^{M^A} \theta_{j,i}^{w-\text{opt}} = 1/M_i^B$ for establishment $j \in U^A$ belonging to enterprise $i \in U^B$, 0 otherwise. This solution agrees with the solution conjectured by Kalton and Brick (1995). They obtained this result based on the simplified situation where $M^A = 2$ and

with s^A obtained through equal probability sampling. Their conclusions suggested the use of optimal values $\theta_{j,i}^{\text{opt}} = 1$ when $\theta_{j,i} > 0$, and $\theta_{j,i}^{\text{opt}} = 0$ when $\theta_{j,i} = 0$. Lavallée (2002) and Lavallée and Caron (2001) obtained results along the same lines by the use of simulations. As mentioned earlier, unfortunately, the weak-optimal weights $\tilde{\theta}_{j,i}^{w-\text{opt}} = 1/M_i^B$ do not correspond to strong-optimal weights that are independent of y.

3 The problem with skewed populations

As mentioned in the introduction, the application of the GWSM to business surveys can produce estimates with large variances. This lack of precision is due to the skewness of the population. We propose to illustrate the problem with a small example given in Figure 3.1.

We want to study the revenue y of the population U^B of Figure 3.1 containing $N^B = 3$ enterprises, where enterprise 1 contains $M_1^B = 4$ establishments, enterprise 2 contains $M_2^B = 4$ establishments, and enterprise 3 contains $M_3^B = 3$ establishments. As it can be observed from Figure 3.1, the revenue y of the $M^B = 11$ establishments can be considered as a skewed population.

For the survey, we build a frame U^A containing the $M^A = 11$ establishments, and we decide to stratify the establishments according to three size strata: stratum $h = 1$ contains the establishments with $y \geq 750$; $h = 2$ contains those with $100 \leq y < 750$; and $h = 3$ those with $y < 100$ (in practice, such a stratification is not possible since the stratification variable y is the same as the variable of interest, and instead, we would use some size variable x highly correlated with the variable of interest y). In stratum $h = 1$, we use a sampling fraction of 1 (*i.e.*, $f_1 = m_1^A / M_1^A = 1$); for $h = 2$, the sample size is 1 (*i.e.*, $f_2 = m_2^A / M_2^A = 1/3$); and for $h = 3$, the sample size is 2 (*i.e.*, $f_3 = m_3^A / M_3^A = 2/6 = 1/3$).

There are $1 \times 3 \times 15 = 45$ possible samples s^A that can be selected from U^A, for estimating the true total $Y = 3{,}800$. For each of these 45 samples, we computed $\hat{Y}$ using (2.1). The estimates are presented in the left box plot of Figure 3.2.

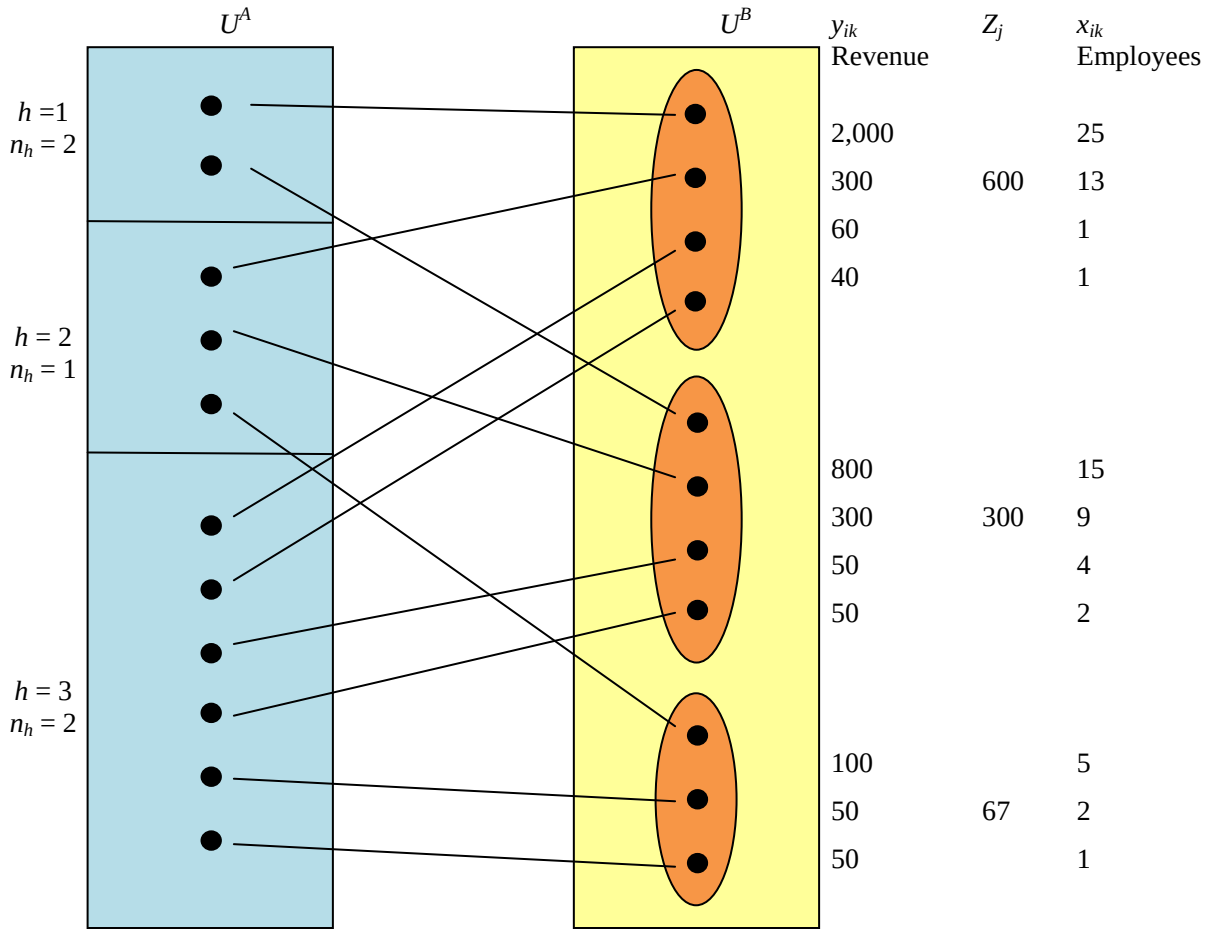


Figure 3.1 Small example

We also computed estimates of Y assuming the use of stratified SRSWoR *without* Indirect Sampling. That is, in each stratum h, we select a sample s_h^A of m_h^A establishments using SRSWoR and we measure only the variable of interest y_{ik} for the establishments ik of U^B directly linked to the sampled establishments j of U^A. Thus, we measure the variable of interest y_j for the sampled establishments j of U^A. Unlike Indirect Sampling, we do not measure the variables of interest of the other establishments of the enterprises containing the initially sampled establishments. This corresponds to the classical sampling theory. Thus, we estimated Y using

$$\hat{Y}_{\text{classic}} = \sum_{h=1}^{H} \frac{M_h^A}{m_h^A} \sum_{j=1}^{m_h^A} y_{hj}. \tag{3.1}$$

It can be proved that estimator (3.1) is unbiased, and its variance is given by

$$\text{Var}(\hat{Y}_{\text{classic}}) = \sum_{h=1}^{H} M_h^A \left(\frac{M_h^A - m_h^A}{m_h^A} \right) S_{y,h}^2 \tag{3.2}$$

where $S_{y,h}^2 = \sum_{j=1}^{M_h^A} \left(y_{hj} - \bar{Y}_h \right)^2 \Big/ (M_h^A - 1)$ and $\bar{Y}_h = \sum_{j=1}^{M_h^A} y_{hj} \Big/ M_h^A$. The estimates are presented in the right box plot of Figure 3.2.

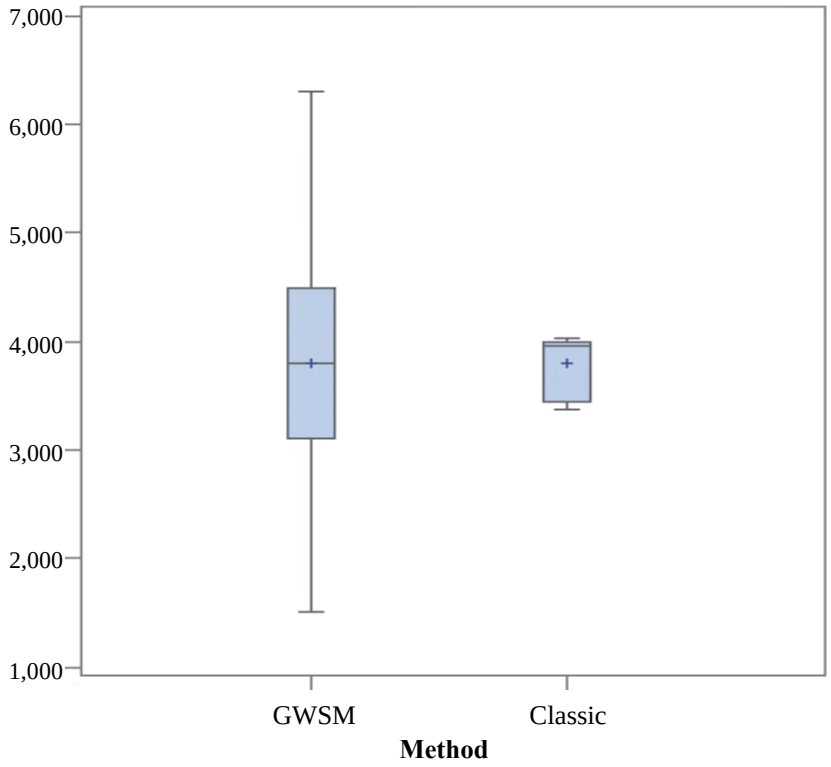


Figure 3.2 Summary of the 45 possible estimates

As we can see from Figure 3.2, the estimates obtained from Indirect Sampling (and the GWSM) are quite variable from one sample to the next. If we do not use Indirect Sampling (*i.e.*, we use the classical approach), the variability is much less. This result can be seen directly from

the variances of $\hat{Y}$ and $\hat{Y}_{\text{classic}}$. Using formulas (2.7) and (3.2), we obtain the variance $V(\hat{Y}_{\text{classic}}) = 80{,}480$, while $V(\hat{Y}) = 1{,}115{,}111$!

The next section presents methods designed to reduce the variability of the estimates produced using Indirect Sampling.

4 Proposed methods

The methods proposed in this section for reducing the variance of the estimates are mainly based on the use of weighted links for the computation of the estimates of Y under Indirect Sampling. We will therefore use estimator (2.9), rather than estimator (2.3). A first set of methods is based on the use of weighted links $\theta_{j,i}$ that are proportional to some measure of size for the establishments. The second set of methods uses the optimal solutions presented in Section 2.2, under different assumptions. Finally, the last set of methods considers the use of the exact selection probabilities, rather than the estimation weights obtained from the GWSM, under two sampling scenarios.

4.1 Methods based on the use of weighted links

Method 1: $\theta_{j,i}$ proportional to π_j^A

We first propose to reduce the variance (2.11) by setting $\theta_{j,i}$ proportional to π_j^A. Formally, this can be written as $\theta_{j,i}^{\pi} = \pi_j^A l_{j,i}$. In business surveys, because stratification is usually done by size (according to some size measure), setting $\theta_{j,i}$ proportional to π_j^A can be viewed as assigning large weights to links of large establishments, and small weights to small ones, which is a natural approach.

With this method, we have $\tilde{\theta}_{j,i}^{\pi} = \theta_{j,i}^{\pi} \Big/ \theta_i^{\pi B} = \pi_j^A l_{j,i} \Big/ \sum_{j=1}^{M^A} \pi_j^A l_{j,i}$. Because of the many-to-one correspondence between U^A and U^B, we obtain $\tilde{\theta}_{j,i}^{\pi} = \theta_{j,i}^{\pi} \Big/ \theta_i^{\pi B} = \pi_j^A l_{j,i} \Big/ \sum_{j=1}^{M_i^B} \pi_j^A$. Therefore, from (2.8), we have

$$w_j^{\pi} = \frac{\sum_{j=1}^{M_i^B} t_j^A}{\sum_{j=1}^{M_i^B} \pi_j^A}. \tag{4.1}$$

Using (4.1), we can rewrite estimator (2.9) as

$$\hat{Y}_{\pi} = \sum_{h=1}^{H} \frac{M_h^A}{m_h^A} \sum_{j=1}^{m_h^A} Z_{hj}^{\pi} \tag{4.2}$$

where

$$Z_{hj}^{\pi} = \frac{\pi_j^A Y_i}{\sum_{j=1}^{M_i^B} \pi_j^A} \tag{4.3}$$

for $j \in h$ and $j \in i$. It should be noted that if all establishments j of a given enterprise belong to the same stratum h, say, we have $\tilde{\theta}_{j,i} = 1 / M_i^B$, and estimator (4.2) is then equivalent to estimator (2.1) (and (2.3)).

For computing the variance of $\hat{Y}_{\pi}$, we use formula (2.11) with the values (4.3). For the example of Section 3, we get $V(\hat{Y}_{\pi}) = 439{,}111$, which is a strong reduction compared to $V(\hat{Y}) = 1{,}115{,}111$, but still relatively far from $V(\hat{Y}_{\text{classic}}) = 80{,}480$.

Method 2: $\theta_{j,i}$ proportional to some size measure x_j

We propose to reduce the variance (2.11) by setting $\theta_{j,i}$ proportional to some size measure x correlated with the variable of interest y. We assume that variable x_j is available for all establishments $j \in U^A$. This variable could be used, for instance, to stratify the sampling frame U^A by size. As for Method 1, setting $\theta_{j,i}^x = x_j l_{j,i}$ can be viewed as assigning large weights to links of large establishments, and small weights to small ones, which again is a natural approach. With this method, we have $\tilde{\theta}_{j,i}^x = \theta_{j,i}^x \Big/ \theta_i^{xB} = x_j l_{j,i} \Big/ \sum_{j=1}^{M^A} x_j l_{j,i}$.

We have $\tilde{\theta}_{j,i}^x = \theta_{j,i}^x \Big/ \theta_i^{xB} = x_j l_{j,i} \Big/ \sum_{j=1}^{M_i^B} x_j = x_j l_{j,i} / X_i$ because of the many-to-one correspondence between U^A and U^B. Therefore, from (2.8), we have

$$w_i^x = \frac{1}{X_i} \sum_{j=1}^{M_i^B} \frac{t_j^A x_j}{\pi_j^A}. \tag{4.4}$$

Using (4.4), we can rewrite (2.9) as

$$\hat{Y}_x = \sum_{h=1}^{H} \frac{M_h^A}{m_h^A} \sum_{j=1}^{m_h^A} Z_{hj}^x \tag{4.5}$$

where

$$Z_{hj}^x = \frac{Y_i}{X_i} x_j \tag{4.6}$$

for $j \in h$ and $j \in i$.

To compute the variance of $\hat{Y}_x$, we use formula (2.11) together with (4.6). For the example of Section 3, the variable x corresponds to the number of employees (see Figure 3.1). The correlation between the revenue y and the number of employees x is relatively high (ρ = 92.8%). We obtain $V(\hat{Y}_x)$ = 686,540, which is again a strong reduction compared to $V(\hat{Y})$ = 1,115,111, but still relatively far from $V(\hat{Y}_{\text{classic}})$ = 80,480.

Method 3: $\theta_{j,i}$ proportional to the variable of interest y_j

The third method proposed is to reduce the variance (2.11) by setting $\theta_{j,i}$ proportional to the variable of interest y measured for the establishment j belonging to enterprise i. Obviously, setting $\theta_{j,i}^y = y_j l_{j,i}$ assigns large weights to links of large establishments, and small weights to small ones, which again is a natural approach. Because y_j is unknown at the beginning of the survey, this method might not look as being implementable since $\theta_{j,i}^y$ depends on y_j. Now, because of the many-to-one correspondence between U^A and U^B, every quantity entering in $\theta_{j,i}^y$ are measured through the Indirect Sampling process.

The proposed method is feasible in this setting and we have $\tilde{\theta}_{j,i}^y = \theta_{j,i}^y \big/ \theta_i^{yB} = y_j l_{j,i} \Big/ \sum_{j=1}^{M_i^B} y_j = y_j l_{j,i} / Y_i$. The weights w_i^y are directly given by (4.4), by replacing x by y. Estimator $\hat{Y}_y$ obtained from (2.9) reduces to

$$\hat{Y}_y = \sum_{h=1}^{H} \frac{M_h^A}{m_h^A} \sum_{j=1}^{m_h^A} y_{hj}, \tag{4.7}$$

which is nothing else than estimator (3.1) obtained from the classical sampling theory.

Note that in general, this method requires one set of weighted links $\theta_{j,i}^{y}$ per variable of interest y. One solution would be to restrict the determination of the weighted links to few key variables of interest, each associated with a larger set of correlated covariates. However, in the present situation, such a restriction is not necessary because estimator (4.7) corresponds simply to estimator (3.1). Indeed, at the end, we obtain estimation weights that simply correspond to the sampling weights.

For computing the variance of (4.7), we simply use formula (3.2). For the example of Section 3, we obtain $V(\hat{Y}_y) = V(\hat{Y}_{\text{classic}}) = 80{,}480$: this is a very large reduction compared to $V(\hat{Y}) = 1{,}115{,}111$.

4.2 Methods using weak-optimal weighted links

Method 4: Using weak-optimal weighted links $\theta_{j,i}^{w-opt,SRS}$ under stratified SRSWoR

This method uses the weak-optimal weighted links $\theta_{j,i}^{w-opt,SRS}$ of Deville and Lavallée (2006) described in Section 2.2. As mentioned earlier, these are obtained by minimising the variance (2.11) for a very specific choice of variable of interest: $Y_i = 1$ for an enterprise i of U^B and $Y_{i'} = 0$ for all other enterprises i' of $U^B (i' \neq i)$. The resulting weak-optimal weighted links do not involve the variable y, per se. Writing the values of $\theta_{j,i}^{w-opt,SRS}$ involves expressions that can be cleverly expressed in matrix notation. Using summations, the expressions become much more complicated, because they involve a mixture of the joint selection probabilities $\pi_{j,j'}^A$ of establishments j and j' that can belong to the same stratum, or not.

Let us define the square matrix $\mathbf{\Delta}^A = \left[\Delta_{j,j'}^A \right]$ of size M^A where $\Delta_{j,j'}^A = (\pi_{j,j'}^A - \pi_j^A \pi_{j'}^A) / \pi_j^A \pi_{j'}^A$. Let $\mathbf{\Gamma}^A = \left[\gamma_{j,j'}^A \right]$ be the inverse of matrix $\mathbf{\Delta}^A$, *i.e.*, $(\mathbf{\Delta}^A)^{-1} = \mathbf{\Gamma}^A$. Let $\mathbf{\Gamma}_i^A = \left[\gamma_{ij,ij'} \right]$ be the square submatrix of $\mathbf{\Gamma}^A$ containing all elements (establishments) (j, j')

belonging to enterprise i. Following Deville and Lavallée (2006), we have $\tilde{\theta}_{j,i}^{w-\text{opt,SRS}} = l_{j,i} \sum_{j'=1}^{M_i^B} \gamma_{ij,ij'} \Big/ \sum_{j=1}^{M_i^B} \sum_{j'=1}^{M_i^B} \gamma_{ij,ij'}$. Unfortunately, for the present case, the many-to-one correspondence between U^A and U^B does not help further in obtaining a simpler form for $\tilde{\theta}_{j,i}^{w-\text{opt,SRS}}$.

Note that if an enterprise i contains an establishment j_0 in the take-all stratum $h = 1$, we have $\Delta_{j_0,j'}^A = (\pi_{j_0,j'}^A - \pi_{j_0}^A \pi_{j'}^A) / \pi_{j_0}^A \pi_{j'}^A = 0$ and the matrix $\mathbf{\Delta}^A$ is singular. In this case, the chosen solution is to set $\tilde{\theta}_{j,i}^{w-\text{opt,SRS}} = 1$ for the take-all establishment j_0 of enterprise i, and $\tilde{\theta}_{j',i}^{w-\text{opt,SRS}} = 0$ for the other establishments $j' \neq j_0$ of enterprise i. This means that $Z_{hj_0}^{w-\text{opt,SRS}} = Y_i$ and $Z_{hj'}^{w-\text{opt,SRS}} = 0$ for $j' \neq j_0$: this means that the complete value Y_i will be assigned to establishment j_0 that contributes 0 to the variance.

We have

$$w_i^{w-\text{opt,SRS}} = \sum_{j=1}^{M_i^B} \frac{t_j^A}{\pi_j^A} \tilde{\theta}_{j,i}^{w-\text{opt,SRS}}. \tag{4.8}$$

Using (4.8), we can rewrite estimator (2.9) as

$$\hat{Y}_{w-\text{opt,SRS}} = \sum_{h=1}^{H} \frac{M_h^A}{m_h^A} \sum_{j=1}^{m_h^A} Z_{hj}^{w-\text{opt,SRS}} \tag{4.9}$$

where

$$Z_{hj}^{w-\text{opt,SRS}} = \sum_{i=1}^{N^B} Y_i \tilde{\theta}_{j,i}^{w-\text{opt,SRS}} \tag{4.10}$$

for $j \in h$ and $j \in i$.

To compute the variance of $\hat{Y}_{w-\text{opt,SRS}}$, we use formula (2.11) with the values (4.10). For the example of Section 3, we get $V(\hat{Y}_{w-\text{opt,SRS}}) = 23{,}111$, which is a tremendous reduction of variance compared to both $V(\hat{Y}) = 1{,}115{,}111$ and $V(\hat{Y}_{\text{classic}}) = 80{,}480$.

Method 5: Using weak-optimal weighted links $\theta_{j,i}^{w-opt,PS}$ under Poisson Sampling

In the context of business surveys, Poisson Sampling selects sample s^A by going through the M^A establishments of population U^A and selecting establishment j if $u_j \leq \pi_j^A$, where $u_j \sim U(0,1)$. The selection probabilities are simply given by $\pi_j^A = m_h^A / M_h^A$ for $j \in h$ and the resulting realised stratum sample size $\tilde{m}_h^A$ is random. In this context, this sampling design can also be seen as stratified Bernoulli Sampling (see Särndal, Swensson and Wretman 1992).

Poisson Sampling (or stratified Bernoulli Sampling) is a very simple sample design. As it can be noted, the selection of each establishment of s^A is done independently from one establishment to another. This means that the joint selection probability $\pi_{j,j'}^A$ of two different establishments j and j' of U^A is simply given by $\pi_{j,j'}^A = \pi_j^A \pi_{j'}^A$. By conditioning on $\tilde{m}_h^A$, it can be shown that stratified Bernoulli Sampling corresponds to stratified SRSWoR. The estimator to be used with stratified Bernoulli Sampling is the ratio estimator

$$\tilde{Y}_\theta = \sum_{h=1}^{H} \frac{M_h^A}{\tilde{m}_h^A} \sum_{j=1}^{\tilde{m}_h^A} Z_{hj}^\theta. \tag{4.11}$$

The variance of estimator (4.11) is approximately given by formula (2.11) (see Brewer and Hanif 1983). Because of the relative closeness between the two designs, assuming Poisson Sampling can be a reasonable approach for computing the weak-optimal weighted links $\theta_{j,i}^{w-opt,PS}$.

The weak-optimal weighted links $\theta_{j,i}^{w-opt,PS}$ are obtained by computing $\tilde{\theta}_{j,i}^{w-opt,SRS}$ as in Method 4, but assuming that sample selection is done using Poisson Sampling. This assumption significantly simplifies the calculations because the matrix $\mathbf{\Delta}^A$ then becomes a diagonal matrix, which is easy to invert. Because of the many-to-one correspondence between U^A and U^B, we obtain after the minimisation process

$$\tilde{\theta}_{j,i}^{w-opt,PS} = \frac{\pi_j^A l_{j,i}}{(1-\pi_j^A)\tau_i} \tag{4.12}$$

where $\tau_i = \sum_{j=1}^{M_i^B} \pi_j^A \big/ (1 - \pi_j^A)$. Therefore, from (2.8), we have

$$w_i^{w-\text{opt,PS}} = \frac{1}{\tau_i} \sum_{j=1}^{M_i^B} \frac{t_j^A}{(1 - \pi_j^A)}. \tag{4.13}$$

Using (4.13), we can rewrite (2.9) as

$$\hat{Y}_{w-\text{opt,PS}} = \sum_{h=1}^{H} \frac{M_h^A}{m_h^A} \sum_{j=1}^{m_h^A} Z_{hj}^{w-\text{opt,PS}} \tag{4.14}$$

where

$$Z_{hj}^{w-\text{opt,PS}} = \frac{\pi_j^A}{(1 - \pi_j^A)} \frac{Y_i}{\tau_i} \tag{4.15}$$

for $j \in h$ and $j \in i$. Note that the previous results assume that $0 < \pi_j^A < 1$ for all establishments j of U^A. For the case where $\pi_j^A = 1$ for a given establishment j_0 of an enterprise i, we set $\tilde{\theta}_{j_0,i}^{w-\text{opt,PS}} = 1$, and $\tilde{\theta}_{j',i}^{w-\text{opt,PS}} = 0$ for $j' \neq j_0$. For computing the variance of $\hat{Y}_{w-\text{opt,PS}}$, we use formula (2.11) with the Z-values given by (4.15).

For the example of Section 3, we get $V(\hat{Y}_{w-\text{opt,PS}}) = 22{,}857$. Again, this is a very large reduction of variance compared to both $V(\hat{Y}) = 1{,}115{,}111$ and $V(\hat{Y}_{\text{classic}}) = 80{,}480$.

Method 6: Using weak-optimal weighted links $\theta_{j,i}^{w-\text{opt,grp}}$ under Poisson Sampling of grouped-establishments

This method consists once more in using the weak-optimal weighted links of Deville and Lavallée (2006) described in Section 2.2, but with grouped-establishments. As a first step, we build grouped-establishments in the population U^A where a grouped-establishment j^* consists in all establishments that are part of the same stratum h and that are belonging to the same enterprise i. This creates a new population U^{A^*} containing M^{A^*} grouped-establishments. The sample s^{A^*} of m^{A^*} grouped-establishments contains all grouped-establishments formed from the

establishments of sample s^A. The selection probability of the grouped-establishment j^* is given by

$$\pi_{j^*}^{A^*} = 1 - \frac{\binom{M_h^A - M_{j^*}}{m_h^A}}{\binom{M_h^A}{m_h^A}} = 1 - \frac{(M_h^A - M_{j^*})(M_h^A - M_{j^*} - 1)...(M_h^A - M_{j^*} - m_h^A + 1)}{M_h^A(M_h^A - 1)...(M_h^A - m_h^A + 1)} \quad (4.16)$$

for $j^* \in h$, where M_{j^*} is the number of establishments within the grouped-establishment j^*.

The rationale behind the use of grouped-establishments is to have only one unit per stratum belonging to a given enterprise. Because, by construction, the grouped-establishments j^* of an enterprise i belong to different strata, their selection is done independently from one grouped-establishment to another. This implies that the solution to weak optimality is then similar to the one obtained in Section 4.5 for Poisson Sampling, but with grouped-establishments. Therefore, we have

$$\tilde{\theta}_{j^*,i}^{w-\text{opt,grp}} = \frac{\pi_{j^*}^A l_{j^*,i}}{(1 - \pi_{j^*}^A)\tau_i^*} \quad (4.17)$$

where $\tau_i^* = \sum_{j^*=1}^{M_i^{B^*}} \pi_{j^*}^{A^*} \Big/ (1 - \pi_{j^*}^{A^*})$ and $M_i^{B^*}$ is the number of groups-establishments contained in enterprise i.

The use of grouped-establishments can be seen as an intermediate step in the Indirect Sampling process going from population U^A to population U^B. That is, the Indirect Sampling process goes from population U^A to population U^{A^*}, and then from population U^{A^*} to population U^B. In the present case, we have $j \in j^* \in i$ for all establishments. Following the rules of transitivity defined by Deville and Lavallée (2006), we can show that the weak-optimal weighted links $\tilde{\theta}_{j,i}^{w-\text{opt,grp}}$ for $j \in j^*$ and $j^* \in i$ (and thus, $j \in i$) are given by

$$\tilde{\theta}_{j,i}^{w-\text{opt,grp}} = \frac{\pi_{j^*}^A l_{j,i}}{(1 - \pi_{j^*}^A)\tau_i^* M_{j^*}}. \quad (4.18)$$

Therefore, from (2.8), we have

$$w_i^{w-\text{opt,grp}} = \frac{1}{\tau_i^*} \sum_{j^*=1}^{M_i^{B^*}} \frac{\pi_{j^*}^{A^*}}{(1-\pi_{j^*}^{A^*})M_{j^*}} \sum_{j=1}^{M_{j^*}} \frac{t_j^A}{\pi_j^A}. \tag{4.19}$$

Using (4.19), we can rewrite (2.9) as

$$\hat{Y}_{w-\text{opt,grp}} = \sum_{h=1}^{H} \frac{M_h^A}{m_h^A} \sum_{j=1}^{m_h^A} Z_{hj}^{w-\text{opt,grp}} \tag{4.20}$$

where

$$Z_{hj}^{w-\text{opt,grp}} = \frac{\pi_{j^*}^{A^*}}{(1-\pi_{j^*}^{A^*})M_{j^*}} \frac{Y_i}{\tau_i^*} \tag{4.21}$$

for $j \in h$ and $j^* \in h$. Note that the previous results assume that $0 < \pi_{j^*}^{A^*} < 1$ for all grouped-establishments j^* of U^{A^*}. For the case where $\pi_{j^*}^{A^*} = 1$ for a given grouped-establishment j_0^* of an enterprise i, we set $\tilde{\theta}_{j,i}^{w-\text{opt,grp}} = l_{j,i} / M_{j^*}$ for the all the establishments j of this grouped-establishment j_0^*, and $\tilde{\theta}_{j,i}^{w-\text{opt,grp}} = 0$ for all other establishments not part of the grouped-establishment j_0^*. We have $\pi_{j^*}^{A^*} = 1$ when at least one establishment j belonging to j^* has $\pi_j^A = 1$. For computing the variance of $\hat{Y}_{w-\text{opt,grp}}$, we use formula (2.11) with the values (4.21).

For the example of Section 3, we get $V(\hat{Y}_{w-\text{opt,grp}}) = 23{,}000$. Again, this is a very significant reduction of variance compared to both $V(\hat{Y}) = 1{,}115{,}111$ and $V(\hat{Y}_{\text{classic}}) = 80{,}480$.

4.3 Other methods

Method 7: Using a designated establishment

As mentioned before, the rationale behind the use of grouped-establishments in Method 6 is to have only one unit per stratum belonging to a given enterprise. Using a similar idea, one can decide on a single establishment that will represent the complete enterprise. That is, for each

enterprise belonging to U^B, we identify one establishment of U^A that will be used for the selection of its owning enterprise. A natural choice for the *designated establishment* within the enterprise is the one with the largest value for a given variable x. For example, x can be the establishment's revenue.

Choosing a designated establishment yields a new sampling frame U^{A_+} that contains the same number of units as the target population U^B, *i.e.*, $M^{A_+} = N^B$. Since there is a one-to-one correspondence between the designated establishment and its owning enterprise, the designated establishment of enterprise i may also be labelled using i. The new frame U^{A_+} can keep the same stratification definition as the original frame U^A. That is, if the stratification of U^A was done by province and industrial classes based on the establishments' values, the stratification of U^{A_+} is done by the same classes based on the designated establishments' values.

From the sampling frame U^{A_+}, we select a sample s^{A_+} of m^{A_+} designated establishments with stratified SRSWoR by using sampling fractions equal to the original ones, *i.e.*, $\pi_i^{A_+} = m_h^{A_+} / M_h^{A_+} = m_h^A / M_h^A$, for $i \in h$. The estimation of the total Y is obtained using the following estimator:

$$\hat{Y}_+ = \sum_{h=1}^{H} \frac{M_h^{A_+}}{m_h^{A_+}} \sum_{i=1}^{m_h^{A_+}} Y_i. \tag{4.22}$$

It can be shown that estimator (4.22) is unbiased, and its variance is given by

$$\mathrm{Var}(\hat{Y}_+) = \sum_{h=1}^{H} M_h^{A_+} \left(\frac{M_h^{A_+} - m_h^{A_+}}{m_h^{A_+}} \right) S_{+Yh}^2 \tag{4.23}$$

where $S_{+Yh}^2 = \sum_{i=1}^{M_h^{A_+}} \left(Y_{hi} - \bar{Y}_h \right)^2 \Big/ (M_h^{A_+} - 1)$ and $\bar{Y}_h = \sum_{i=1}^{M_h^{A_+}} Y_{hi} \Big/ M_h^{A_+}$.

Note that although we only keep one designated establishment per enterprise, we are still able to produce estimates per stratum, or for any domain of interest (*e.g.*, different industrial activities). For example, let us consider the small example of Section 3. For the first enterprise of

U^B (*i.e.*, the one with a total revenue of 2,400), the designated establishment would be the first establishment of the take-all stratum of U^A (*i.e.*, the establishment with 25 employees). None of the three other establishments of this enterprise would be available for sampling. However, if we were interested in producing an estimate for the second stratum, we would simply restrict the computation of the values Y_i in (4.22) to the establishments belonging to this second stratum. In the present case, rather than using $Y_i = 2{,}400$ in (4.22), we would then use $Y_i = 300$. This corresponds to domain estimation (Särndal, *et al.* 1992).

For the example of Section 3, we obtain $V(\hat{Y}_+)$ = 1,820,000! With this method, since an establishment inherits all revenues of the enterprise, the use of a designated establishment is advantageous when this establishment is in the take-all stratum. However, the designated establishment may itself be contained in a take-some stratum, and this results in a stratum that is even more skewed. The total revenue of the enterprise, multiplied by the sampling weight, is assigned to this take-some stratum, and this increases the variance significantly.

Method 8: Using the selection probabilities of the enterprises

As mentioned in Lavallée (2002, 2007), using the Rao-Blackwell theorem, sufficient statistics can improve an existing estimator by producing a new estimator with a mean squared error that is smaller than or equal to that of the initial estimator (see Cassel, Särndal and Wretman 1977). Note that this form of improvement was used, for instance, by Thompson (1990) in the context of Adaptive Cluster Sampling.

Starting from estimator (2.1) (or (2.3)), the estimator $\hat{Y}_{\mathrm{RB}}$ obtained by using the Rao-Blackwell theorem is given by

$$\hat{Y}_{\mathrm{RB}} = \sum_{i=1}^{n^B} \frac{Y_i}{L_i^B} \sum_{j=1}^{M^A} \frac{P(t_j^A = 1 \mid s^B)}{\pi_j^A} l_{j,i} \tag{4.24}$$

where $P(t_j^A = 1 \mid s^B)$ is the probability of having selected establishment j from U^A, given that the n^B enterprises of s^B have been selected from U^B.

Using the many-to-one correspondence between U^A and U^B, an approximation to the probability $P(t_j^A = 1 \mid s^B)$ can be obtained. That is, for $j \in i$, we have

$$\begin{aligned} P(t_j^A = 1 \mid s^B) &\approx P(t_j^A = 1 \mid i \in s^B) \\ &= P(t_j^A = 1, i \in s^B) \,/\, P(i \in s^B) \\ &= P(t_j^A = 1) \,/\, P(i \in s^B) \\ &= \pi_j^A \,/\, \pi_i^B \end{aligned} \tag{4.25}$$

where π_i^B is the selection probability of enterprise $i \in U^B$, which corresponds to the probability of selecting any of its M_i^B establishments. Note that result (4.25) becomes exact in the context of Poisson Sampling. Using (4.25), estimator (4.24) is then approximately equivalent to the following Horvitz-Thompson estimator

$$\hat{Y}_{\mathrm{HT}} = \sum_{i=1}^{n^B} \frac{Y_i}{\pi_i^B}. \tag{4.26}$$

Since estimator (4.26) is nothing else than a Horvitz-Thompson estimator based on the selection of enterprises, its variance is given by

$$\hat{Y}_{\mathrm{HT}} = \sum_{i=1}^{N^B} \sum_{i'=1}^{N^B} \frac{(\pi_{i,i'}^B - \pi_i^B \pi_{i'}^B)}{\pi_i^B \pi_{i'}^B} Y_i Y_{i'}. \tag{4.27}$$

The computation of the selection probability π_i^B requires the knowledge of the selection probabilities π_j^A of all M_i^B establishments of enterprise i. In general, this can be difficult or even impossible to obtain (see Lavallée 2002, 2007). This can be a severe barrier for using estimator (4.26) in practice, and actually, this is one of the driving reason why using the GWSM. However, in the present case, this reveals to be possible because the complete frame U^A is available for the selection of the establishments. The task is also simplified by the use of stratified SRSWoR. The selection probabilities π_i^B can then be computed by adapting formula (4.16). It is also possible to compute the joint selection probabilities $\pi_{i,i'}^B$, but this is more difficult.

For the example of Section 3, we obtain $V(\hat{Y}_{\mathrm{HT}}) = 14{,}545$, and this value corresponds to the lowest variance of the proposed methods.

5 Simulations using real data

The simulations reflect a typical business survey at Statistics Canada. Three populations commonly surveyed by Statistics Canada were chosen. The populations of establishments of the industries of Manufacturing, Retail Trade and Restaurants were extracted from the Business Register (BR). These populations are known to have a skewed distribution for economic variables such as revenue, especially the first two. Stratified SRSWoR was used for the simulations, stratified by the industry, the region and the class of revenue. The algorithm of Lavallée-Hidiroglou (1988) was used to create the classes of revenue, determine the sample size and perform the allocation. The establishments were divided in three strata based on their size: one take-all and two take-some strata. A coefficient of variation of 5% was targeted in each of the strata by industry and region. The following table contains some statistics on the population.

Table 5.1
Simulation populations, sample counts and statistics

Industry	N^B	M^A	m^A	Average revenue	Variance	Skewness
Manufacturing	96,955	100,109	2,223	4,364,808	1.08×10^{16}	164
Retail Trade	142,020	159,247	3,627	2,034,111	3.29×10^{14}	133
Restaurants	107,358	113,425	2,439	561,764	4.43×10^{12}	106
Total	346,333	372,781	8,289		---	

The revenue variable available on the BR was used as the variable of interest y. For these simulations, since the values of this variable are known for all units, no sample selection was

needed for most of the methods. It should be noted that for Method 2 ($\theta_{j,i}$ proportional to some size measure x_j), the number of employees was used.

Except for Methods 7 and 8, the true variances were calculated from the data using formula (2.11). For Method 7, we used formula (4.23). For Method 8, we needed to calculate the true probabilities of selection π_i^B of all enterprises. Although it would have been possible to compute these probabilities with exact formulas, we choose to compute them using a two-step Monte-Carlo simulation. One reason for this is variance formula (4.27) that uses the joint selection probabilities $\pi_{i,i'}^B$ which involves too many pairs (j, j'). For the first step of the Monte-Carlo simulation, we selected 20,000 samples of establishments using the stratified SRSWoR design described above. For each sample, we identified which enterprise ended up being selected. Over those 20,000 samples, we were able to estimate the probability of selection π_i^B of each enterprise i under that design of unequal probabilities. Once these probabilities were derived, we conducted another Monte-Carlo simulation for computing the variance. We selected $R = 20{,}000$ samples s^A of establishments using again the stratified SRSWoR design described above. We then obtained the corresponding samples s^B of enterprises. For each replicate $r, r = 1, \ldots, R$, we produced an estimate $\hat{Y}_{\mathrm{HT},r}$ of Y using estimator (4.26). The variance was computed using

$$V_{\mathrm{MC}}(\hat{Y}_{\mathrm{HT}}) = \frac{1}{R}\sum_{r=1}^{R}(\hat{Y}_{\mathrm{HT},r} - \hat{\bar{Y}}_{\mathrm{HT}}^{(R)})^2 \tag{5.1}$$

where $\hat{\bar{Y}}_{\mathrm{HT}}^{(R)} = \sum_{r=1}^{R}\hat{Y}_{\mathrm{HT},r} / R$. Note that because estimator (4.26) is unbiased, we have $\hat{\bar{Y}}_{\mathrm{HT}}^{(R)} \approx Y$.

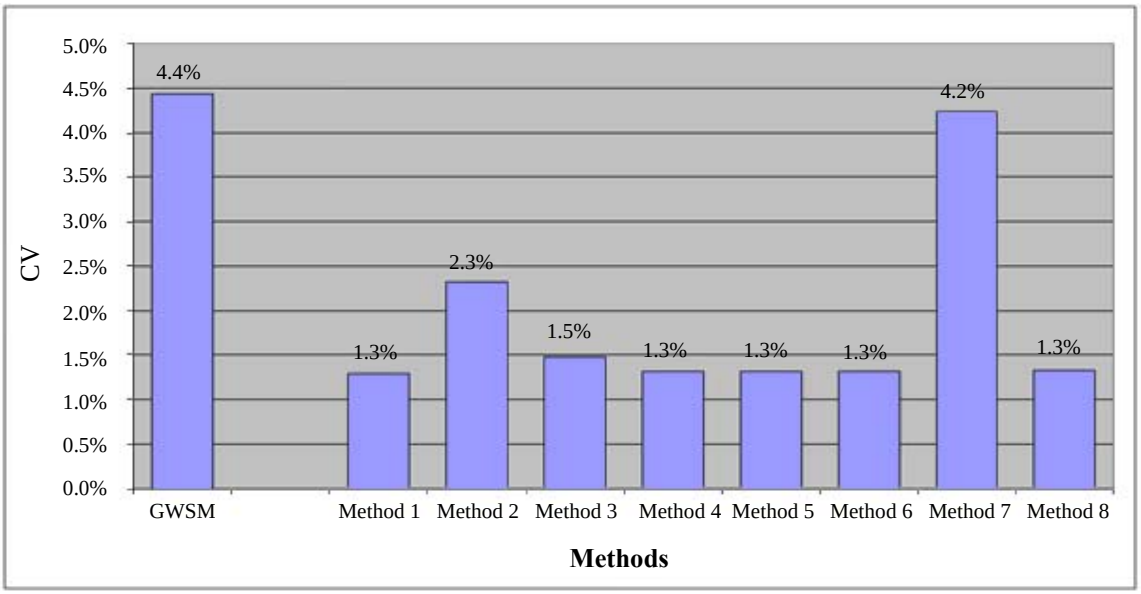
For each estimator $\hat{Y}$, the coefficient of variation was computed using

$$\mathrm{CV}(\hat{Y}) = \frac{\sqrt{V(\hat{Y})}}{Y}. \tag{5.2}$$

5.1 Results of the simulation

For the classical GWSM and all the methods presented, the estimates, variances and coefficients of variation were computed. The graph below presents the CVs obtained at the national level.

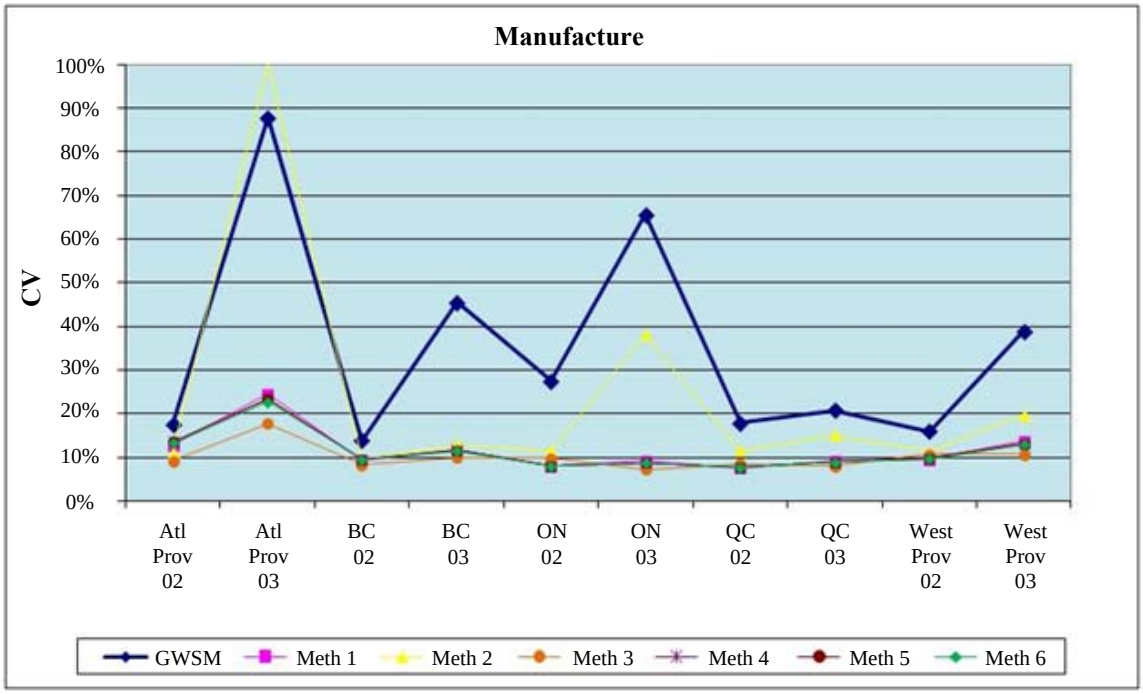
Graph 5.1 Coefficients of variation by method



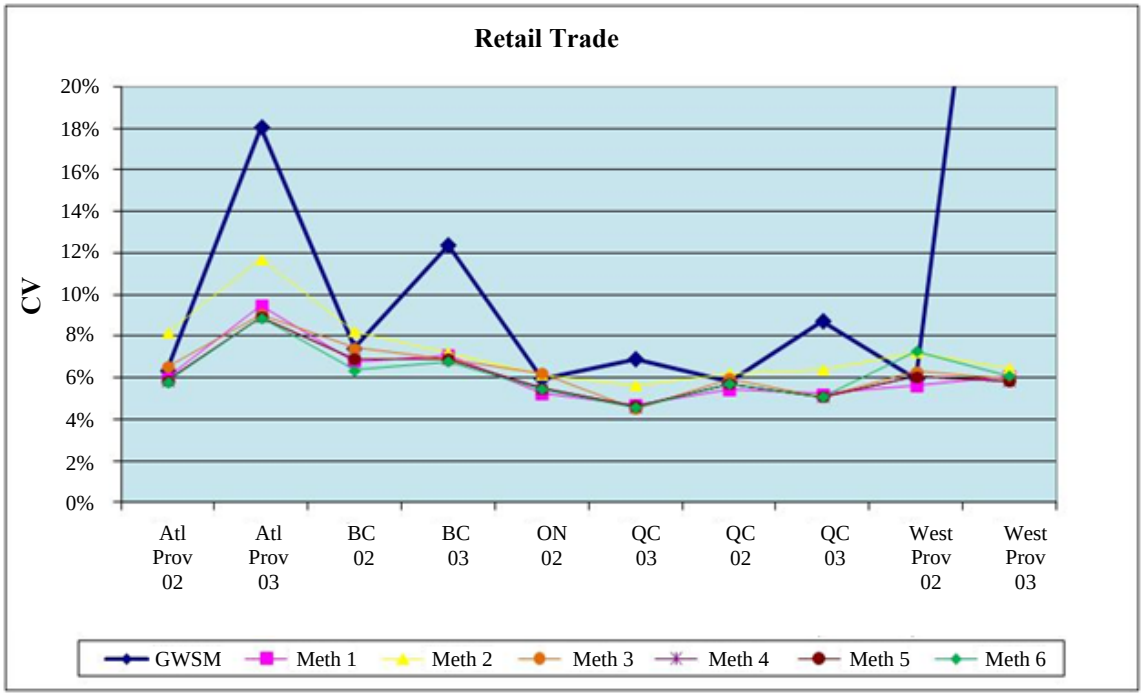
Except for Method 7, all methods show a decrease in the variance, and often the reduction is substantial. As described earlier, Method 7 (using designated establishments) determines a single establishment within an enterprise based on the auxiliary variable, and the whole enterprise is assigned to that establishment. In other words, a given establishment inherits all the revenue of the enterprise. This is beneficial when the designated establishment is in a take-all stratum. However, if the designated establishment is in a take-some stratum, the distribution within that stratum will become more skewed. One hundred percent of the revenue of the enterprise, times its sampling weight is assigned in that stratum, and the resulting variance is increased significantly. All the other methods provided reasonable results and we analyze them in detail via the graphs that follow.

The following graphs show the CV for each take some strata by industry.

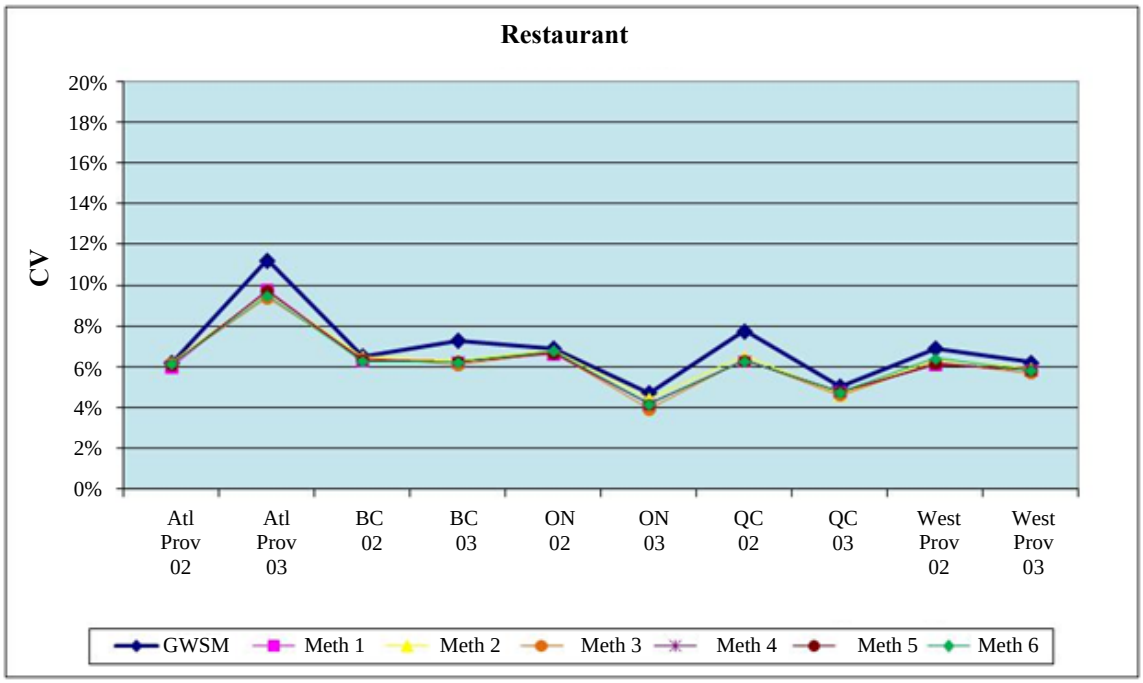
Graph 5.2 CV for Manufacturing by strata



Graph 5.3 CV for Retail Trade by Strata



Graph 5.4 CV for restaurants by strata



Note 1 The scale of the CV is different for Manufacturing than for the two other industries.

Note 2 The CVs per stratum of methods 7 and 8 are not showed here because they are not pertinent for the present comparison. The reason is that for these two methods, the notion of stratum is not the same as for the other methods. The stratification defined by the original sample design is at the establishment level. For methods 7 and 8, sampling was done at the enterprise level, and therefore a typical stratum for methods 1 to 6 became a domain for methods 7 and 8. Of course, the variances associated with methods 7 and 8 are much higher, making any comparison with the other methods irrelevant.

The CVs are particularly high for the classical GWSM in some strata, especially in the industry of Manufacturing. This was expected because the skewness of the distribution of the variable of interest was the highest among the three industries. Furthermore, in this industry, we have establishments with revenues that can vary a lot within the same enterprise, and these establishments can be spread amongst several strata. It is for these reasons that the variance of the classical GWSM ends up being very high.

All graphs show that there is a reduction in the CV, thus in the variance, by using any of the suggested methods. The CV are generally lower for the other methods, as compared to the classical GWSM (represented by the dark blue line with diamonds).

5.2 Comparison of the proposed methods

Method 1 yields very promising results, given its simplicity. This method results in some of the lowest CV amongst all methods. It is really targeting the source of the problem of the GWSM: the need of an unequal allocation of the weights, somewhat proportional to the size of the variable of interest. Since most of business surveys are using stratification by size, this method works well. It also has the advantage of not depending directly on the variable of interest.

Method 2 uses an auxiliary variable (here, the number of employees) to distribute the weights. This variable is not that well correlated with the variable of interest, and this explains why it has the lowest decrease in variance. In fact, this method is a weaker version of Method 3.

Method 3 is sharing the weight proportionally to the variable of interest y, which is the establishment revenue within the enterprise. The method performs very well both at the national and provincial level. It is generally slightly higher than method 1, 4, 5 and 6 because of the high skewness of the distribution of the revenue.

Methods 4, 5 and 6 give very similar results, offering a CV between 6% and 10% for Retail Trade and Restaurants, and between 10% and 25% for Manufacturing. The similarity in the results for all three methods is reasonable because they all aim to produce the lowest possible variance. Whenever there is one establishment of an enterprise that is in a take-all stratum, these methods concentrate all values on this establishment, and assign links of zero to all other establishments of the enterprise. This is a natural choice to minimize the variance since the contribution to the variance of that enterprise becomes null. Since an establishment of a large enterprise can be part of a take-all stratum, the variance turns out to be lower than for any other methods. It is for these reasons that these three methods are the best way to share the weight. However, it was not possible to determine which one of the three was the best.

Method 7 is not providing good results. Recall that for this method, we let a single designated establishment represent the whole enterprise. All the establishment values of the variable of interest are then summed to the enterprise, and assigned to that designated establishment. The

distribution of the variable of interest by stratum becomes even more skewed and this leads to a larger variance. From a sampling viewpoint, an enterprise ends up in a single stratum (because it is represented by a single establishment), which might not be take-all. In addition, this is not efficient for producing estimates at the provincial or industry level. The estimates cannot benefit from the stratification of establishments, while this is possible for the other methods. This also contributes to a higher variance.

Method 8 is using the selection probabilities of the enterprises obtained via simulations. This method is doing well with a CV of 1.3% at the national level. It matches Method 4, 5 and 6. However, this method can reveal to be difficult to apply in practice. We must either calculate explicitly the second order selection probabilities $\pi_{i,i'}^{B}$, which can be very difficult to obtain, or estimate them by simulation, which is very time consuming.

6 Conclusion

The GWSM can be used in the context of business surveys, but it can result in large variances because of the skewness of the data associated with such populations. Eight alternate methods to share the weights were proposed to reduce the variance. A simulation was conducted to compare those using real data. The simulations replicated a typical business survey using stratified SRSWoR. All proposed methods showed a reduction in variance. The simulations showed that the best methods were obtained by sharing the weights proportional to the π_j^A (Method 1), or using weak-optimal weighted links (Methods 4, 5 and 6). The weighted links under Poisson sampling (Method 5) is the preferred method. In theory, it is close to the optimal method, and it is also the simplest to implement.

References

Brewer, K.R.W., and Hanif, M. (1983). *Sampling With Unequal Probabilities*. New York: Springer-Verlag, 164 pages.

Cassel, C.-M., Särndal, C.-E. and Wretman, J.H. (1977). *Foundations of Inference in Survey Sampling*. New York: John Wiley & Sons, Inc.

Delorme, C. (2000). Data allocation: A new survey process for business surveys. ICES-II, Proceedings of the Second International Conference on Establishment Surveys, Survey Methods for Businesses, Farms, and Institutions, Buffalo, New York, June 17-21, 2000, 1285-1290.

Deville, J.-C., and Lavallée, P. (2006). Indirect sampling: The foundations of the generalized weight share method. *Survey Methodology*, Vol. 32, 2, 165-176.

Kalton, G., and Brick, J.M. (1995). Weighting schemes for household panel surveys. *Survey Methodology*, Vol. 21, 1, 33-44.

Lavallée, P. (2002). *Le Sondage Indirect, ou la Méthode Généralisée du Partage des Poids*. Éditions de l'Université de Bruxelles, Brussels.

Lavallée, P. (2007). *Indirect Sampling*. New York: Springer.

Lavallée, P., and Caron, P. (2001). Estimation using the generalised weight share methods: The case of record linkage. *Survey Methodology*, Vol. 27, 2, 155-169.

Lavallée, P., and Hidiroglou, M.A. (1988). On the stratification of skewed populations. *Survey Methodology*, Vol. 14, 1, 33-43.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Statistics Canada (2010). *A Brief Guide to the BR*. Business Register Division, Statistics Canada, Ottawa, 9 pages, July 2010.

Thompson, S.K. (1990). Adaptive cluster sampling. *Journal of the American Statistical Association*, Vol. 85, 412, 1050-1059.

Thompson, S.K. (1991). Stratified adaptive cluster sampling. *Biometrika*, Vol. 78, 2, 389-397.

Thompson, S.K. (1992). *Sampling*. New York: John Wiley & Sons, Inc.

Thompson, S.K. (2002). *Sampling, 2nd Edition*. New York: John Wiley & Sons, Inc.

Thompson, S.K., and Seber, G.A. (1996). *Adaptive Sampling*. New York: John Wiley & Sons, Inc.

Wolter, K.M. (2007). Introduction to Variance Estimation, 2nd Edition. New York: Springer.


ELECTRONIC PUBLICATIONS AVAILABLE AT
PUBLICATIONS ÉLECTRONIQUES DISPONIBLE À
www.statcan.gc.ca

Survey Methodology, June 2013
Vol. 39, No. 1, pp. 217-229
Statistics Canada, Catalogue No. 12-001-X

On the performance of self benchmarked small area estimators under the Fay-Herriot area level model

Yong You, J.N.K. Rao and Mike Hidiroglou[1]

Abstract

We consider two different self-benchmarking methods for the estimation of small area means based on the Fay-Herriot (FH) area level model: the method of You and Rao (2002) applied to the FH model and the method of Wang, Fuller and Qu (2008) based on augmented models. We derive an estimator of the mean squared prediction error (MSPE) of the You-Rao (YR) estimator of a small area mean that, under the true model, is correct to second-order terms. We report the results of a simulation study on the relative bias of the MSPE estimator of the YR estimator and the MSPE estimator of the Wang, Fuller and Qu (WFQ) estimator obtained under an augmented model. We also study the MSPE and the estimators of MSPE for the YR and WFQ estimators obtained under a misspecified model.

Key Words: Augmented model; Empirical best linear unbiased prediction; Mean squared prediction error; Model misspecification.

1 Introduction

Subpopulations or domains are called small areas when the domain sample sizes are not large enough to provide reliable area-specific direct estimates of domain parameters. In those cases, it becomes necessary to use model-based indirect estimators that make use of sample data from related areas through linking models to achieve significant gain in efficiency over direct estimators. In this paper we focus on a basic area level model, called the Fay-Herriot (FH) model (Fay and Herriot 1979), and associated empirical best linear unbiased predictors (EBLUPs) of small area means.

Those EBLUPs do not necessarily agree with direct estimators for aggregates which are preferred in practice because they are based on large enough sample sizes to satisfy reliability requirements and do not depend on models. As a result, benchmarking is often done to force agreement. In this paper, we focus on two methods of self benchmarking, based on modifications

1. Yong You, Statistical Research and Innovation Division, Statistics Canada, Ottawa, Canada. E-mail: yong.you@statcan.gc.ca; J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, Canada; Mike Hidiroglou, Statistical Research and Innovation Division, Statistics Canada, Ottawa, Canada.

to the EBLUPs. The first method (WFQ), due to Wang, Fuller and Qu (Wang *et al.* 2008), uses an augmented FH model and the associated EBLUPs, denoted as WFQ estimators. The second method (YR), also proposed by Wang, Fuller and Qu, is based on an approach used by You and Rao (2002) in the context of unit level models. The YR estimators are obtained by modifying the optimal estimators of the regression parameters used in the EBLUPs to force agreement. Because of the modifications to the EBLUPs, benchmarked estimators WFQ and YR will have higher mean squared predication errors (MSPEs).

This paper has two main objectives. The first objective is to compare MSPEs and their estimators for the YR and WFQ estimators of small area means. Section 2 reviews the expressions for MSPEs and associated estimators for the EBLUP and the WFQ estimators. Section 3 develops expressions for MSPE and its estimator for the YR estimator. Section 4 compares the MSPEs and their estimators in a simulation study.

The second objective of our paper is to examine the performance of the WFQ and YR estimators and their MSPE estimators when the linking model in the FH model is misspecified due to an omitted variable. WFQ also studied the effect of misspecification of the linking model for a particular example, for which they showed that the YR estimator leads to large bias whereas the WFQ estimator did not. However, this result was due to the fact that in their simulation study, the augmenting variable was highly correlated with an omitted covariate (correlation coefficient $\rho = 0.983$). As a result, the augmented model used was in fact close to the true model, leading to the superior performance for the WFQ estimator. In Section 4, we consider a different omitted variable that is weakly correlated with the augmenting variable ($\rho = -0.116$) so that the augmented model is also misspecified. We compare the biases, MSPEs and their estimators for the EBLUP, YR and WFQ estimators obtained under the misspecified model.

2 EBLUPs and WFQ estimators

Suppose that we have m small areas with design-unbiased direct estimators, y_i, of the area means $\theta_i, i = 1, \ldots, m$. The FH model refers to (y_i, θ_i) and associated area level auxiliary variables $x_i = (x_{i1}, \ldots, x_{ip})'$ with $x_{i1} = 1$. Assuming independent sampling across areas, the FH model may be written as a linear mixed model given by

$$y_i = \theta_i + e_i = x_i'\beta + v_i + e_i, i = 1, \ldots, m, \tag{2.1}$$

where $\theta_i = x_i'\beta + v_i$ is the linking model, and e_i is the sampling error with mean 0 and known variance σ_i^2, which is independent of the area specific random effect v_i. Sampling is independent across areas and the v_i's are assumed to be independent and identically distributed with mean 0 and variance σ_v^2.

The best linear unbiased predictor (BLUP) of θ_i under the "true" model (2.1) is given by

$$\tilde{\theta}_i = \gamma_i y_i + (1 - \gamma_i)x_i'\tilde{\beta}, \tag{2.2}$$

where $\gamma_i = \sigma_v^2 / (\sigma_v^2 + \sigma_i^2)$ and $\tilde{\beta}$ is the optimal weighted least squared (WLS) estimator of β given by

$$\tilde{\beta} = \left(\sum_{i=1}^{m} \gamma_i x_i x_i' \right)^{-1} \left(\sum_{i=1}^{m} \gamma_i x_i y_i \right), \tag{2.3}$$

see Rao (2003, page 116). The estimator $\tilde{\theta}_i$ depends on the unknown model variance σ_v^2, and replacing σ_v^2 in (2.2) by a suitable estimator $\hat{\sigma}_v^2$, we get the EBLUP:

$$\hat{\theta}_i^{\text{EBLUP}} = \hat{\gamma}_i y_i + (1 - \hat{\gamma}_i)x_i'\hat{\beta}, \tag{2.4}$$

where $\hat{\gamma}_i = \hat{\sigma}_v^2 / (\hat{\sigma}_v^2 + \sigma_i^2)$ and $\hat{\beta}$ is obtained from (2.3) by replacing γ_i by $\hat{\gamma}_i$. In this paper, we use the restricted maximum likelihood (REML) estimator of σ_v^2, assuming normality of v_i and e_i. The weighted sum $\sum_{i=1}^{m} w_i \hat{\theta}_i^{\text{EBLUP}}$ of the EBLUPs (2.4) does not necessarily agree with the corresponding weighted direct estimator $\sum_{i=1}^{m} w_i y_i$ of the aggregate, where the w_i's are

pre-specified weights such that $\sum_{i=1}^{m} w_i y_i$ is a design-consistent estimator of the aggregate (total or mean). If the gap between $\sum_{i=1}^{m} w_i \hat{\theta}_i^{\text{EBLUP}}$ and $\sum_{i=1}^{m} w_i y_i$ is large, it may indicate some model failure that should be taken care of before proceeding to benchmarking, as noted by the Associate Editor.

An estimator of the mean squared prediction error $\text{MSPE}(\hat{\theta}_i^{\text{EBLUP}}) = \text{E}(\hat{\theta}_i^{\text{EBLUP}} - \theta_i)^2$ correct to second-order terms, under REML estimation, is given by

$$\text{mspe}(\hat{\theta}_i^{\text{EBLUP}}) = g_{1i} + g_{2i} + 2g_{3i}, \tag{2.5}$$

where $g_{1i} = \hat{\gamma}_i \sigma_i^2$ is the leading term of order $O(1)$, and g_{2i} and g_{3i} are lower order terms of order $O(m^{-1})$ accounting for the variability of $\hat{\beta}$ and $\hat{\sigma}_v^2$ respectively (Rao 2003, page 128). We have

$$g_{2i} = (1 - \hat{\gamma}_i)^2 x_i' \hat{V}(\tilde{\beta}) x_i = \hat{\sigma}_v^2 (1 - \hat{\gamma}_i)^2 x_i' \left(\sum_{i=1}^{m} \hat{\gamma}_i x_i x_i' \right)^{-1} x_i$$

and

$$g_{3i} = 2(\sigma_i^2)^2 (\hat{\sigma}_v^2 + \sigma_i^2)^{-3} \left\{ \sum_{i=1}^{m} (\hat{\sigma}_v^2 + \sigma_i^2)^{-2} \right\}^{-1},$$

where $\hat{V}(\tilde{\beta})$ is the estimator of $\text{V}(\tilde{\beta}) = \sigma_v^2 \left(\sum_{i=1}^{m} \gamma_i x_i x_i' \right)^{-1}$. The MSPE estimator (2.5) is nearly unbiased in the sense that

$$\text{E}\{\text{mspe}(\hat{\theta}_i^{\text{EBLUP}})\} = \text{MSPE}(\hat{\theta}_i^{\text{EBLUP}}) + O(m^{-2}).$$

WFQ obtained an EBLUP estimator, $\hat{\eta}_i^{\text{EBLUP}}$, under the following augmented FH model:

$$y_i = x_i' \delta + w_{ei} \lambda + u_i + e_i \equiv \eta_i + e_i, \tag{2.6}$$

where the random effects u_i are independent with $\text{E}(u_i) = 0$ and $\text{Var}(u_i) = \sigma_u^2$, and is independent of e_i. The augmenting auxiliary variable is taken as $w_{ei} = w_i \sigma_i^2$. WFQ showed that

the EBLUP estimator of η_i under the augmented model (2.6), $\hat{\eta}_i^{\text{EBLUP}} \equiv \hat{\theta}_i^{\text{WFQ}}$, is self-benchmarking in the sense of satisfying

$$\sum_{i=1}^{m} w_i \hat{\theta}_i^{\text{WFQ}} = \sum_{i=1}^{m} w_i y_i .$$

The EBLUP $\hat{\theta}_i^{\text{WFQ}}$ under the augmented model (2.6) is obtained from (2.4) by changing x_i' to (x_i', w_{ei}), $\hat{\sigma}_v^2$ to $\hat{\sigma}_u^2$ and $\hat{\beta}'$ to $(\hat{\delta}', \hat{\lambda})$. The estimator of MSPE$(\hat{\theta}_i^{\text{WFQ}})$ is similarly obtained from (2.5). Under the true model (2.1), the MSPE of $\hat{\theta}_i^{\text{WFQ}}$ will be larger than the MSPE of $\hat{\theta}_i^{\text{EBLUP}}$, but its estimator, mspe$(\hat{\theta}_i^{\text{WFQ}})$, will remain nearly unbiased under the true model, as noted by the Associate Editor and a referee, because the true model is a special case of the augmented model with $\lambda = 0$.

3 YR estimator

WFQ applied the You and Rao (2002) method to model (2.1) and obtained an estimator of θ_i given by

$$\hat{\theta}_i^{\text{YR}} = \hat{\gamma}_i y_i + (1 - \hat{\gamma}_i) x_i' \hat{\beta}^{\text{YR}}, \tag{3.1}$$

where $\hat{\beta}^{\text{YR}}$ is obtained from

$$\tilde{\beta}^{\text{YR}} = \left\{ \sum_{i=1}^{m} w_i (1 - \gamma_i) x_i x_i' \right\}^{-1} \left\{ \sum_{i=1}^{m} w_i (1 - \gamma_i) x_i y_i \right\} \tag{3.2}$$

by replacing γ_i by $\hat{\gamma}_i$. Note that the YR estimator (3.1) has the same form as the EBLUP $\hat{\theta}_i^{\text{EBLUP}}$, but it uses a non-optimal estimator for β. The YR estimators $\hat{\theta}_i^{\text{YR}}$ are self-benchmarking, *i.e.*, $\sum_{i=1}^{m} w_i \hat{\theta}_i^{\text{YR}} = \sum_{i=1}^{m} w_i y_i$, since by (3.2)

$$\sum_{i=1}^{m} w_i (1 - \gamma_i)(y_i - x_i' \tilde{\beta}^{\text{YR}}) = 0.$$

However, the MSPE of $\hat{\theta}_i^{\mathrm{YR}}$ will be slightly higher than the MSPE of $\hat{\theta}_i^{\mathrm{EBLUP}}$ based on $\hat{\beta}$, because $\hat{\beta}$ is asymptotically more efficient than $\hat{\beta}^{\mathrm{YR}}$.

As in the case of $\hat{\theta}_i^{\mathrm{EBLUP}}$, the estimator of $\mathrm{MSPE}(\hat{\theta}_i^{\mathrm{YR}})$ has g_{1i}, g_{2i} and g_{3i} terms. We need to estimate the variance of $\tilde{\beta}^{\mathrm{YR}}$ in order to get the g_{2i} term in the estimator of $\mathrm{MSPE}(\hat{\theta}_i^{\mathrm{YR}})$; the other terms g_{1i} and g_{3i} are not affected. It follows from (3.2) that

$$\mathrm{V}(\tilde{\beta}^{\mathrm{YR}}) = \sigma_v^2 \left\{ \sum_{i=1}^{m} w_i(1-\gamma_i)x_i x_i' \right\}^{-1} \left\{ \sum_{i=1}^{m} w_i^2(1-\gamma_i)^2 \gamma_i^{-1} x_i x_i' \right\} \left\{ \sum_{i=1}^{m} w_i(1-\gamma_i)x_i x_i' \right\}^{-1}. \quad (3.3)$$

The estimator $\hat{V}(\tilde{\beta}^{\mathrm{YR}})$ is obtained by substituting $\hat{\sigma}_v^2$ and $\hat{\gamma}_i$ for σ_v^2 and γ_i in (3.3).

The estimator of $\mathrm{MSPE}(\hat{\theta}_i^{\mathrm{YR}})$ is given by

$$\mathrm{mspe}(\hat{\theta}_i^{\mathrm{YR}}) = g_{1i} + g_{2i}^{\mathrm{YR}} + 2g_{3i}, \quad (3.4)$$

where

$$g_{2i}^{\mathrm{YR}} = (1-\hat{\gamma}_i)^2 x_i' \hat{V}(\tilde{\beta}^{\mathrm{YR}}) x_i.$$

The MSPE estimator (3.4) is nearly unbiased under the true model (2.1), similar to the MSPE estimator (2.5) of $\hat{\theta}_i^{\mathrm{EBLUP}}$.

Remark: Any estimator $\hat{y}_i$ of θ_i may be adjusted as

$$\hat{y}_i^a = \hat{y}_i + a_i \left(\sum_{i=1}^{m} w_i y_i - \sum_{i=1}^{m} w_i \hat{y}_i \right)$$

for specified a_i to satisfy the benchmarking constraint $\sum_{i=1}^{m} w_i \hat{y}_i^a = \sum_{i=1}^{m} w_i y_i$, where $\sum_{i=1}^{m} w_i a_i = 1$. In particular, we can use $\hat{y}_i = \hat{\theta}_i^{\mathrm{EBLUP}}$ to obtain the adjusted EBLUP estimator. As noted by WFQ, both $\hat{\theta}_i^{\mathrm{YR}}$ and $\hat{\theta}_i^{\mathrm{WFQ}}$ are estimators of the form $\hat{y}_i^a$ because $\sum_{i=1}^{m} w_i y_i - \sum_{i=1}^{m} w_i \hat{y}_i$ is equal to zero when $\hat{y}_i$ is set equal to $\hat{\theta}_i^{\mathrm{YR}}$ or $\hat{\theta}_i^{\mathrm{WFQ}}$. Any set of estimators $\{\hat{y}_i\}$ that satisfy $\sum_{i=1}^{m} w_i y_i = \sum_{i=1}^{m} w_i \hat{y}_i$ has the self-benchmarking property.

4 Simulation study

We conducted a small simulation study on the bias and MSPE of YR and WFQ estimators and on the relative bias and coefficient of variation (CV) of YR and WFQ estimators of MSPE, using the simulation setup of WFQ. As in WFQ, we considered $m = 50$ small areas (states) with area sample sizes n_i ranging from 7 to 58, and known sampling variances $\sigma_i^2 = 16 / n_i$. The weights w_i used in the benchmarking constraint are taken as $w_i = \left(\sum_{i=1}^{50} P_i \right)^{-1} P_i$, where P_i is the population of state i reported in Table 4.1 of WFQ.

In WFQ, the true model is given by

$$y_i = \beta_0 + \beta_1 z_i + v_i + e_i, \tag{4.1}$$

with $\beta_0 = 6$, $\beta_1 = 3$, $v_i \sim \mathrm{N}(0, \sigma_v^2 = 1)$ and $e_i \sim \mathrm{N}(0, 16 / n_i)$. WFQ used $z_i = P_i^{0.2} - \left(\sum_{i=1}^{50} P_i^{0.2} \right) \Big/ 50$ as the covariate values in the true model, and the chosen z_i's are highly correlated with the associated weighted augmented variable $w_{ei} = w_i \sigma_i^2$; correlation coefficient $\rho = 0.983$.

In WFQ, the mis-specified model is taken as the mean model

$$y_i = \mu + \tilde{v}_i + e_i \tag{4.2}$$

with $\tilde{v}_i \sim \mathrm{N}(0, \tilde{\sigma}_v^2)$ and $e_i \sim \mathrm{N}(0, 16 / n_i)$, and this choice makes the associated augmented model

$$y_i = \delta_0 + \lambda w_i \sigma_i^2 + u_i + e_i \tag{4.3}$$

almost the same as the true model (4.1).

To reflect misspecification by the augmented model, we generated 50 values from $\mathrm{N}(1, 1)$ and these 50 values $(z_1, \ldots, z_{50})$ were held fixed across the simulations. The resulting correlation coefficient between the generated z_i and $w_i \sigma_i^2$ is -0.116. We then took the corresponding model

$$y_i = \beta_0 + \beta_1 z_i + v_i + e_i \tag{4.4}$$

with $\beta_0 = 6$, $\beta_1 = 3$, $v_i \sim \mathrm{N}(0,1)$ and $e_i \sim \mathrm{N}(0, 16/n_i)$ as our true model and the mean model (4.2) as the mis-specified model. We considered two scenarios: (i) True model is (4.4) and it is correctly modeled and specified. (ii) True model is (4.4) and it is incorrectly modeled and specified as the mean model (4.2).

We generated $R = 10{,}000$ data sets $\{(y_i^{(r)}, z_i), i = 1, ..., 50; r = 1, \ldots, R\}$ under the true model (4.4), where $y_i^{(r)} = \beta_0 + \beta_1 z_i + v_i^{(r)} + e_i^{(r)}$ with $v_i^{(r)}$ generated from $\mathrm{N}(0,1)$ and $e_i^{(r)}$ from $\mathrm{N}(0, 16/n_i)$, respectively. Let $\theta_i^{(r)} = \beta_0 + \beta_1 z_i + v_i^{(r)}$ denote the value of the true mean θ_i and $\hat{\theta}_i^{(r)}$ the associated value of an estimator $\hat{\theta}_i$ (EBLUP, YR or WFQ) for the r^{th} simulation run under scenario (i) or scenario (ii). Then the simulated absolute bias and MSPE of $\hat{\theta}_i$ are given by

$$\left|\mathrm{B}(\hat{\theta}_i)\right| = \left| R^{-1}\sum_{r=1}^{R}(\hat{\theta}_i^{(r)} - \theta_i^{(r)})\right|$$

and

$$\mathrm{MSPE}(\hat{\theta}_i) = R^{-1}\sum_{r=1}^{R}(\hat{\theta}_i^{(r)} - \theta_i^{(r)})^2.$$

The absolute relative bias (ARB) of an MSPE estimator, $\mathrm{mspe}(\hat{\theta}_i)$, is calculated as $\left|\mathrm{RB}\{\mathrm{mspe}(\hat{\theta}_i)\}\right|$, where

$$\mathrm{RB}\{\mathrm{mspe}(\hat{\theta}_i)\} = \frac{1}{R}\sum_{r=1}^{R}\{\mathrm{mspe}(\hat{\theta}_i^{(r)}) - \mathrm{MSPE}(\hat{\theta}_i)\} \Big/ \mathrm{MSPE}(\hat{\theta}_i),$$

where $\mathrm{mspe}(\hat{\theta}_i^{(r)})$ is the value of $\mathrm{mspe}(\hat{\theta}_i)$ for the r^{th} simulation run. Similarly, the coefficient of variation (CV) of $\mathrm{mspe}(\hat{\theta}_i)$ is calculated as

$$\mathrm{CV}\{\mathrm{mspe}(\hat{\theta}_i)\} = \left[\frac{1}{R}\sum_{r=1}^{R}\{\mathrm{mspe}(\hat{\theta}_i^{(r)}) - \mathrm{MSPE}(\hat{\theta}_i)\}^2\right]^{\frac{1}{2}} \Big/ \mathrm{MSPE}(\hat{\theta}_i).$$

We summarize the simulation results by reporting the first quartile, third quartile, median and mean of the values of $\left|\mathrm{B}(\hat{\theta}_i)\right|$, $\mathrm{MSPE}(\hat{\theta}_i)$, $\left|\mathrm{RB}\{\mathrm{mspe}(\hat{\theta}_i)\}\right|$%, and $\mathrm{CV}\{\mathrm{mspe}(\hat{\theta}_i)\}$%. We also grouped the areas by the sampling variances, $\sigma_i^2 = 16/n_i$, into three groups and then calculated the

above measures separately for each group. Results obtained are very similar to those without grouping. Therefore, we report results only for the ungrouped case, for simplicity.

Table 4.1 reports the summary measures for the absolute bias of the estimators EBLUP, YR and WFQ, under scenario (i) and scenario (ii). As expected, the values of the absolute bias measures are all negligible for the three estimators under scenario (i), confirming their unbiasedness under scenario (i). On the other hand, the absolute bias is large relative to standard error under scenario (ii) for all the three estimators, leading to much higher MSPE relative to scenario (i). It is interesting to note from the values under scenario (ii) in Table 4.1 that the absolute bias is virtually the same across the estimators, suggesting that self-benchmarking may not reduce the absolute bias under model misspecification. A plausible explanation is that the YR estimator attaches the same weight, $\hat{\gamma}_i$, to the direct estimator; similarly, the WFQ estimator attaches a weight close to $\hat{\gamma}_i$. Our result does not contradict the conclusion of WFQ that the WFQ estimator reduces the absolute bias, because their augmented model virtually eliminated the model misspecification.

Table 4.1
Summary measures for absolute bias of EBLUP, YR and WFQ estimators

Measure	EBLUP	YR	WFQ
	Scenario (i)		
1st quartile	0.002	0.002	0.002
Median	0.005	0.005	0.005
Mean	0.007	0.007	0.007
3rd quartile	0.009	0.010	0.009
	Scenario (ii)		
1st quartile	0.092	0.085	0.083
Median	0.144	0.153	0.152
Mean	0.254	0.255	0.251
3rd quartile	0.316	0.311	0.314

It may also be noted that the adjusted EBLUP estimator $\hat{y}_i^a$ defined in the Remark using $\hat{y}_i = \hat{\theta}_i^{\text{EBLUP}}$ might perform better than the YR and WFQ estimators in terms of absolute bias under Scenario (ii). For example, if the EBLUP estimator is underestimating in most of the areas, then the correction term $a_i \left(\sum_{i=1}^m w_i y_i - \sum_{i=1}^m w_i \theta_i^{\text{EBLUP}} \right)$ will be positive and hence compensate for the underestimation. On the other hand, the correction term is zero for the self-benchmarking estimators YR and WFQ.

Table 4.2 reports the summary measures for MSPE of the estimators EBLUP, YR and WFQ, under scenarios (i) and (ii). It shows that, under scenario (i), MSPEs of the three estimators EBLUP, YR and WFQ are practically the same, except that EBLUP has slightly smaller MSPE as expected. Under scenario (ii), we again observe very little difference in MSPEs of the three estimators. However, the MSPE of a given estimator is considerably increased due to model misspecification. For example, the third quartile (corresponding to small n_i) increased from 0.621 to 1.100 for EBLUP, 0.626 to 1.094 for YR and 0.634 to 1.117 for WFQ. This inflation in MSPE is due to bias induced by model misspecification.

Table 4.2
Summary measures for MSPE of EBLUP, YR and WFQ estimators

Measure	EBLUP	YR	WFQ
	Scenario (i)		
1st quartile	0.403	0.406	0.403
Median	0.471	0.476	0.472
Mean	0.499	0.505	0.507
3rd quartile	0.621	0.626	0.634
	Scenario (ii)		
1st quartile	0.562	0.567	0.567
Median	0.751	0.748	0.752
Mean	0.885	0.886	0.883
3rd quartile	1.100	1.094	1.117

Table 4.3 reports the summary measures for percentage absolute RB and percentage CV of the MSPE estimators associated with EBLUP, YR and WFQ, under scenario (i) corresponding to no misspecification of the true model (4.4). Table 4.4 reports the results on the MSPE estimators under scenario (ii), corresponding to misspecification of the true model (4.4) as the mean model (4.2).

Table 4.3
Summary measures for absolute relative bias (%) and CV (%) of MSPE estimators associated with EBLUP, YR and WFQ under scenario (i): Model (4.4) correctly specified

Measure	EBLUP	YR	WFQ
Absolute Relative Bias (ARB) %			
1st quartile	0.678	0.650	0.627
Median	1.186	1.216	1.173
Mean	1.354	1.349	1.372
3rd quartile	1.674	1.716	1.863
Coefficient of variation (CV) %			
1st quartile	12.93	12.44	13.19
Median	15.64	15.40	15.78
Mean	16.02	15.69	15.78
3rd quartile	19.39	19.07	19.05

Table 4.4
Summary measures for absolute relative bias (%) and CV (%) of MSPE estimators associated with EBLUP, YR and WFQ under scenario (ii): Model (4.4) mis-specified

Measure	EBLUP	YR	WFQ
Absolute Relative Bias (ARB) %			
1st quartile	4.848	4.850	4.857
Median	7.653	7.776	7.466
Mean	10.412	10.469	10.221
3rd quartile	13.423	13.611	13.776
Coefficient of variation (CV) %			
1st quartile	5.510	5.579	4.700
Median	7.894	7.317	7.610
Mean	10.52	10.55	10.17
3rd quartile	13.42	13.03	13.11

Turning to the summary measures of absolute relative bias (ARB) and CV of MSPE estimators, Table 4.3 shows that under scenario (i), ARB% is very small (< 2%) for all the three MSPE estimators, confirming that the MSPE estimators are nearly unbiased. Also, CV% is practically the same for all the three MSPE estimators. Under scenario (ii), Table 4.4 shows that ARB% is significantly increased, but practically the same for all the three MSPE estimators. For example, the third quartile increased from 1.674 to 13.423 in the case of EBLUP, 1.716 to 13.611 in the case of YR and 1.863 to 13.776 in the case of WFQ. Also, CV% is practically the same for all the three MSPE estimators. Comparing the CV values in Table 4.4 (Scenario (ii)) with the corresponding CV values in Table 4.3 (Scenario (i)), we see that the CV's are actually much lower under Scenario (ii) than under Scenario (i). This is because the MSPE (used in the denominator of the CV measure) is much higher under Scenario (ii) than under Scenario (i).

5 Concluding remarks

We have studied the properties of the EBLUP and two self-benchmarking estimators, YR and WFQ, under the Fay-Herriot area level model. We presented a nearly unbiased estimator of MSPE of the YR estimator in Section 3. Our simulation results indicate that the three methods perform very similarly with respect to MSPE of the estimators and RB and CV of the MSPE estimators. However, under gross model misspecification due to an omitted variable, MSPE is significantly increased for all the three estimators, unless the augmented model of WFQ nearly eliminates misspecification. In practice, it is difficult to account for model misspecification due to unknown omitted variables. It is also interesting to note that self-benchmarking may not reduce bias under model misspecification that is not accounted by the augmented model of WFQ.

Acknowledgements

We are thankful to a referee and the Associate Editor for many constructive comments and suggestions.

References

Fay, R.E., and Herriot, R.A. (1979). Estimation of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 268-277.

Rao, J.N.K. (2003). *Small Area Estimation*, New York: John Wiley & Sons, Inc.

Wang, J., Fuller, W.A. and Qu, Y. (2008). Small area estimation under a restriction. *Survey Methodology*, 34, 1, 29-36.

You, Y., and Rao, J.N.K. (2002). A pseudo empirical best linear unbiased prediction approach to small area estimation using survey weights. *The Canadian Journal of Statistics*, 30, 431-439.


ELECTRONIC PUBLICATIONS AVAILABLE AT
PUBLICATIONS ÉLECTRONIQUES DISPONIBLE À
www.statcan.gc.ca

Survey Methodology, June 2013
Vol. 39, No. 1, pp. 231-241
Statistics Canada, Catalogue No. 12-001-X

Conservative variance estimation for sampling designs with zero pairwise inclusion probabilities

Peter M. Aronow and Cyrus Samii[1]

Abstract

We consider conservative variance estimation for the Horvitz-Thompson estimator of a population total in sampling designs with zero pairwise inclusion probabilities, known as "non-measurable" designs. We decompose the standard Horvitz-Thompson variance estimator under such designs and characterize the bias precisely. We develop a bias correction that is guaranteed to be weakly conservative (nonnegatively biased) regardless of the nature of the non-measurability. The analysis sheds light on conditions under which the standard Horvitz-Thompson variance estimator performs well despite non-measurability and where the conservative bias correction may outperform commonly-used approximations.

Key Words: Horvitz-Thompson estimation; Non-measurable designs; Variance estimation.

1 Introduction

Sampling designs sometimes result in pairs of units having zero probability of being jointly included in the sample. Horvitz and Thompson (1952)'s statement of the properties of the finite population total makes clear that general, unbiased variance estimation for estimators of population totals is impossible for such *non-measurable* designs (Särndal, Swensson and Wretman 1992, page 33). Optimal methods for variance estimation in these cases remains an open problem. This paper analyzes the nature of the biases that non-measurability introduces for the standard Horvitz-Thompson estimator and studies an approach to correct for this bias in a conservative manner. While our results cannot offer a solution to the non-measurability problem for all practical applications, we do clarify conditions under which the standard estimator performs well and where the conservative bias correction outperforms commonly-used approximations.

1. Peter M. Aronow, Department of Political Science, Yale University, 77 Prospect St., New Haven, CT 06520. E-mail: peter.aronow@yale.edu; Cyrus Samii, Department of Politics, New York University, 19 West 4th St., New York, NY 10012. E-mail: cds2083@nyu.edu.

Despite their theoretical drawbacks, sampling designs with zero pairwise inclusion probabilities are quite common. A common non-measurable design is one that draws only single units or clusters from a set of strata. This may occur if the population or a subpopulation of interest is incidentally sparse over stratification cells. Another common non-measurable design is a systematic sample in which unit indices are sampled from a list in multiples from a random starting value. In these designs, units whose indices are multiples from different starting values have zero joint probability of inclusion.

Approximate methods have been proposed for special cases, as discussed in Hansen, Hurwitz and Madow (1953, Section 9.15), Särndal *et al.* (1992, Chapter 3), and Wolter (2007, Chapters 2 and 8). In the single-unit per stratum case, a common approach is to collapse strata and assume units were drawn via a simple random sample from the larger, collapsed stratum. For systematic samples, the standard approach is to use an approximation based on an assumption of simple random sampling with replacement. These approximate methods are generally biased to a degree that cannot be determined from the data. In some cases, it can be shown that the bias will tend to be positive, but such is not the case generally, and especially so when the zero pairwise inclusion probabilities occur in a haphazard manner.

This paper begins in Section 2 by decomposing the bias of the Horvitz-Thompson variance estimator under non-measurability. This exposes precisely how conditions on the underlying data result in more or less bias. In Section 3, we also show how a simple application of Young's inequality yields a bias correction and a class of estimators guaranteed to have weakly positive bias as well as no bias under special conditions. We discuss implications for applied work in Section 4.

2 Variance estimation for the Horvitz-Thompson estimator

Consider a population U indexed by $1, \ldots, k, \ldots N$ and a sampling design such that the probability of inclusion in the sample for unit k is π_k, and the joint inclusion probability for

units k and l is π_{kl}. Under a measurable design, there are two conditions: (1) $\pi_k > 0$ and π_k is known for all $k \in U$ and (2) $\pi_{kl} > 0$ and π_{kl} is known for all $k, l \in U$. Non-measurable designs include those for which either of the two conditions for a measurable design do not hold. Failure to meet the former condition precludes unbiased estimation of totals.

The Horvitz-Thompson estimator of a population total is $\hat{t} = \sum_{k \in s} y_k / \pi_k = \sum_{k \in U} I_k \, y_k / \pi_k$, where $I_k \in \{0, 1\}$ is unit k's inclusion indicator, the only stochastic component of the expression, with $\mathrm{E}(I_k) = \pi_k$, the inclusion probability, and s and U refer to the sample and the population, respectively. Define $\mathrm{E}(I_k I_l) = \pi_{kl}$, which is the probability that both units k and l from U are included in the sample. Since $I_k I_k = I_k$, $\mathrm{E}(I_k I_k) = \pi_{kk} = \pi_k$ by construction. When condition 1 holds, as we assume throughout, the Horvitz-Thompson estimator is unbiased.

2.1 Properties of the Horvitz-Thompson variance estimator under measurability

By Horvitz and Thompson (1952), the variance of the Horvitz-Thompson estimator for the total is

$$\mathrm{Var}(\hat{t}) = \sum_{k \in U} \sum_{l \in U} \mathrm{Cov}(I_k, I_l) \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} = \sum_{k \in U} \mathrm{Var}(I_k) \left(\frac{y_k}{\pi_k} \right)^2 + \sum_{k \in U} \sum_{l \in U \setminus k} \mathrm{Cov}(I_k, I_l) \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}.$$

We label a sample from a measurable design, s^M, and an unbiased estimator for $\mathrm{Var}(\hat{t})$ on s^M is

$$\widehat{\mathrm{Var}}(\hat{t}) = \sum_{k \in s^M} \sum_{l \in s^M} \frac{\mathrm{Cov}(I_k, I_l)}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} = \sum_{k \in U} \sum_{l \in U} I_k I_l \frac{\mathrm{Cov}(I_k, I_l)}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l},$$

where the only stochastic part of the expression is $I_k I_l$, and unbiasedness is by $\mathrm{E}(I_k I_l) = \pi_{kl}$.

2.2 Properties of the Horvitz-Thompson variance estimator under non-measurability

We now examine the case where condition 2 does not hold: $\pi_{kl} = 0$ for some units $k, l \in U$. Because I_k is a Bernoulli random variable with probability π_k, $\mathrm{Cov}(I_k, I_l) = \pi_{kl} - \pi_k \pi_l$ for

$k \neq l$, and $\text{Cov}(I_k, I_k) = \text{Var}(I_k) = \pi_k(1 - \pi_k)$. Then, we can re-express the variance above as,

$$\text{Var}(\hat{t}) = \sum_{k \in U} \pi_k(1 - \pi_k) \left(\frac{y_k}{\pi_k} \right)^2 + \sum_{k \in U} \sum_{l \in U \setminus k} (\pi_{kl} - \pi_k \pi_l) \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$$

$$= \sum_{k \in U} \pi_k(1 - \pi_k) \left(\frac{y_k}{\pi_k} \right)^2 + \sum_{k \in U} \sum_{l \in \{U \setminus k : \pi_{kl} > 0\}} (\pi_{kl} - \pi_k \pi_l) \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} - \underbrace{\sum_{k \in U} \sum_{l \in \{U \setminus k : \pi_{kl} = 0\}} y_k y_l}_{A}.$$

For k and l such that $\pi_{kl} = 0$, the sampling design will never permit unbiased estimation of the component of the variance labeled as A above, since we will never observe y_k and y_l together. We label a sample from a design where condition 2 fails as s^0. When $\widehat{\text{Var}}(\hat{t})$ is applied to s^0, the result is unbiased for $\text{Var}(\hat{t}) + A$. We state this formally as follows:

Proposition 1. When s^0 refers to a sample from a design with some $\pi_{kl} = 0$,

$$\text{E}\left[\widehat{\text{Var}}(\hat{t}) \right] = \text{Var}(\hat{t}) + \sum_{k \in U} \sum_{l \in \{U \setminus k : \pi_{kl} = 0\}} y_k y_l = \text{Var}(\hat{t}) + A.$$

Proof. The result follows from,

$$\text{E}\left[\sum_{k \in s^0} \sum_{l \in s^0} \frac{\text{Cov}(I_k, I_l)}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \right] = \text{E}\left[\sum_{k \in U} \sum_{l \in \{U : \pi_{kl} > 0\}} I_k I_l \frac{\text{Cov}(I_k, I_l)}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \right]$$

$$= \sum_{k \in U} \text{Var}(I_k) \left(\frac{y_k}{\pi_k} \right)^2 + \sum_{k \in U} \sum_{l \in \{U \setminus k : \pi_{kl} > 0\}} \text{Cov}(I_k, I_l) \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$$

$$= \text{Var}(\hat{t}) + \sum_{k \in U} \sum_{l \in \{U \setminus k : \pi_{kl} = 0\}} y_k y_l = \text{Var}(\hat{t}) + A.$$

The standard Horvitz-Thompson variance estimator, if applied to designs with zero pairwise inclusion probabilities, can therefore have a positive or a negative bias. If the y_k, y_l values are always nonnegative (or always nonpositive), then the bias is always nonnegative. When values may be positive or negative, A is the sum of cross-products of outcomes that never appear

together under the design sample. If the jointly exclusive outcomes are centered over zero, then no correlation in these outcomes would tend to result in small bias, positive correlation in positive bias, and negative correlation in negative bias.

3 Conservative bias correction for the Horvitz-Thompson variance estimator under non-measurability

The case where A may be less than zero for a non-measurable design suggests the need for some adjustment that will guarantee a bias that is weakly bounded below by zero. We first develop a general bias correction that is guaranteed to be conservative, later providing a special simplified case for practical usage.

3.1 General formulation

Consider the following variance estimator:

$$\widehat{\mathrm{Var}}_C(\hat{t}) = \sum_{k\in U}\sum_{l\in\{U:\pi_{kl}>0\}} I_k I_l \frac{\mathrm{Cov}(I_k, I_l)}{\pi_{kl}} \frac{y_k}{\pi_k}\frac{y_l}{\pi_l} + \sum_{k\in U}\sum_{l\in\{U\setminus k:\pi_{kl}=0\}} \left(I_k \frac{|y_k|^{a_{kl}}}{a_{kl}\pi_k} + I_l \frac{|y_l|^{b_{kl}}}{b_{kl}\pi_l} \right),$$

where a_{kl}, b_{kl} are positive real numbers such that $1/a_{kl} + 1/b_{kl} = 1$ for all pairs k, l with $\pi_{kl} = 0$. The estimator is guaranteed to produce an expected value greater than or equal to the true variance for all designs, and is thus conservative. We state this property formally:

Proposition 2. The expected value of $\widehat{\mathrm{Var}}_C(\hat{t})$,

$$\mathrm{E}\left[\widehat{\mathrm{Var}}_C(\hat{t})\right] \geq \mathrm{Var}(\hat{t}).$$

Proof. By Young's inequality,

$$\frac{|y_k|^{a_{kl}}}{a_{kl}} + \frac{|y_l|^{b_{kl}}}{b_{kl}} \geq |y_k||y_l|,$$

if $1/a_{kl} + 1/b_{kl} = 1$. Define A^* such that,

$$A^* = \sum_{k \in U} \sum_{l \in \{U \setminus k : \pi_{kl} = 0\}} \frac{\left| y_k \right|^{a_{kl}}}{a_{kl}} + \frac{\left| y_l \right|^{b_{kl}}}{b_{kl}} \geq \sum_{k \in U} \sum_{l \in \{U \setminus k : \pi_{kl} = 0\}} \left| y_k \right| \left| y_l \right| \geq \sum_{k \in U} \sum_{l \in \{U \setminus k : \pi_{kl} = 0\}} y_k y_l = A$$

and

$$A^* \geq \sum_{k \in U} \sum_{l \in \{U \setminus k : \pi_{kl} = 0\}} \left| y_k \right| \left| y_l \right| \geq \sum_{k \in U} \sum_{l \in \{U \setminus k : \pi_{kl} = 0\}} -y_k y_l = -A.$$

Therefore

$$\text{Var}(\hat{t}) + A + A^* \geq \text{Var}(\hat{t}).$$

The associated Horvitz-Thompson estimator of A^* would be

$$\hat{A}^* = \sum_{k \in U} \sum_{l \in \{U \setminus k : \pi_{kl} = 0\}} \left(I_k \frac{\left| y_k \right|^{a_{kl}}}{a_{kl} \pi_k} + I_l \frac{\left| y_l \right|^{b_{kl}}}{b_{kl} \pi_l} \right),$$

which is unbiased by $E(I_k) = \pi_k$ and $E(I_l) = \pi_l$.

Since $E(\hat{A}^*) = A^*$, by Proposition 1,

$$\text{E}\left[\sum_{k \in s^0} \sum_{l \in s^0} \frac{\text{Cov}(I_k, I_l)}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} + \hat{A}^* \right] = \text{Var}(\hat{t}) + A + A^*$$

$$\text{E}\left[\widehat{\text{Var}}_C(\hat{t}) \right] \geq \text{Var}(\hat{t}).$$

Substituting terms,

$$\text{E}\left[\sum_{k \in U} \sum_{l \in \{U : \pi_{kl} > 0\}} I_k I_l \frac{\text{Cov}(I_k, I_l)}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} + \sum_{k \in U} \sum_{l \in \{U \setminus k : \pi_{kl} = 0\}} \left(I_k \frac{\left| y_k \right|^{a_{kl}}}{a_{kl} \pi_k} + I_l \frac{\left| y_l \right|^{b_{kl}}}{b_{kl} \pi_l} \right) \right] \geq \text{Var}(\hat{t}).$$

$\widehat{\text{Var}}_C(\hat{t})$ is justified as a conservative estimator for the case when A is not known to be positive. This estimator is unbiased under a special condition:

Corollary 1. If, for all pairs k, l such that $\pi_{kl} = 0$, (i) $\left| y_k \right|^{a_{kl}} = \left| y_l \right|^{b_{kl}}$ and (ii) $-y_k y_l = \left| y_k \right| \left| y_l \right|$,

$$\mathrm{E}\left[\widehat{\mathrm{Var}}_C(\hat{t})\right] = \mathrm{Var}(\hat{t}).$$

Proof. By (i), (ii) and Young's inequality,

$$\frac{|y_k|^{a_{kl}}}{a_{kl}} + \frac{|y_l|^{b_{kl}}}{b_{kl}} = |y_k||y_l| = -y_k y_l.$$

Therefore,

$$A^* = \sum_{k\in U}\sum_{l\in\{U\setminus k:\pi_{kl}=0\}} \frac{|y_k|^{a_{kl}}}{a_{kl}} + \frac{|y_l|^{b_{kl}}}{b_{kl}} = \sum_{k\in U}\sum_{l\in\{U\setminus k:\pi_{kl}=0\}} |y_k||y_l| = \sum_{k\in U}\sum_{l\in\{U\setminus k:\pi_{kl}=0\}} -y_k y_l = -A.$$

It follows that $\mathrm{Var}(\hat{t}) + A + A^* = \mathrm{Var}(\hat{t})$ and $\mathrm{E}\left[\widehat{\mathrm{Var}}_C(\hat{t})\right] = \mathrm{Var}(\hat{t})$.

If any units k, l are in clusters (*i.e.*, $\Pr(I_k \neq I_l) = 0$), these units should be totaled into one larger unit before estimation. Combining units will tend to reduce the bias of the variance estimator because only pairs of cluster-level totals will be included in A^*, as opposed to all constituent pairs.

3.2 Simplified special case

In general, it would be difficult to assign optimal values of a_{kl} and b_{kl} for all pairs k, l such that $\pi_{kl} = 0$. Instead, we examine one intuitive case, assigning all $a_{kl} = b_{kl} = 2$:

$$\widehat{\mathrm{Var}}_{C2}(\hat{t}) = \sum_{k\in U}\sum_{l\in\{U:\pi_{kl}>0\}} I_k I_l \frac{\mathrm{Cov}(I_k, I_l)}{\pi_{kl}} \frac{y_k}{\pi_k}\frac{y_l}{\pi_l} + \sum_{k\in U}\sum_{l\in\{U\setminus k:\pi_{kl}=0\}} \left(I_k \frac{y_k^2}{2\pi_k} + I_l \frac{y_l^2}{2\pi_l} \right).$$

As a special case of $\widehat{\mathrm{Var}}_C(\hat{t})$, $\widehat{\mathrm{Var}}_{C2}(\hat{t})$ is also conservative:

Corollary 2. The expected value of $\widehat{\mathrm{Var}}_{C2}(\hat{t})$,

$$\mathrm{E}\left[\widehat{\mathrm{Var}}_{C2}(\hat{t})\right] \geq \mathrm{Var}(\hat{t}).$$

Proof. For all pairs k, l such that $\pi_{kl} = 0, 1/a_{kl} + 1/b_{kl} = 1$. Proposition 1 therefore holds.

The choice to set all $a_{kl} = b_{kl} = 2$ is justified by the fact that it will yield the lowest value of the estimator $\widehat{\text{Var}}_C(\hat{t})$ subject to the constraint that a_{kl} and b_{kl} are fixed as constants a and b over all k, l.

Corollary 3. Among the class of estimators $\widehat{\text{Var}}_{\text{Cab}}(\hat{t})$, defined as the set of estimators $\widehat{\text{Var}}_C(\hat{t})$ such that all $a_{kl} = a$ and all $b_{kl} = b$, $\widehat{\text{Var}}_{C2}(\hat{t}) = \min_{a,b}\left[\widehat{\text{Var}}_{\text{Cab}}(\hat{t})\right]$.

Proof. By simple algebra,

$$\widehat{\text{Var}}_{\text{Cab}}(\hat{t}) = \sum_{k\in U}\sum_{l\in\{U:\pi_{kl}>0\}} I_k I_l \frac{\text{Cov}(I_k, I_l)}{\pi_{kl}} \frac{y_k}{\pi_k}\frac{y_l}{\pi_l} + \sum_{k\in U}\sum_{l\in\{U\setminus k:\pi_{kl}=0\}}\left(I_k \frac{|y_k|^a}{a\pi_k} + I_l \frac{|y_l|^b}{b\pi_l}\right)$$

$$= \widehat{\text{Var}}(\hat{t}) + \sum_{k\in U}\sum_{l\in\{U\setminus k:\pi_{kl}=0\}}\left(I_k \frac{|y_k|^a}{a\pi_k} + I_l \frac{|y_l|^b}{b\pi_l}\right)$$

$$= \widehat{\text{Var}}(\hat{t}) + \frac{1}{2}\sum_{k\in U}\sum_{l\in\{U\setminus k:\pi_{kl}=0\}}\left(I_k \frac{|y_k|^a}{a\pi_k} + I_l \frac{|y_l|^b}{b\pi_l} + I_k \frac{|y_k|^b}{b\pi_k} + I_l \frac{|y_l|^a}{a\pi_l}\right)$$

$$= \widehat{\text{Var}}(\hat{t}) + \frac{1}{2}\sum_{k\in U}\sum_{l\in\{U\setminus k:\pi_{kl}=0\}}\left(\frac{I_k}{\pi_k}\left(\frac{|y_k|^a}{a} + \frac{|y_k|^b}{b}\right) + \frac{I_l}{\pi_l}\left(\frac{|y_l|^a}{a} + \frac{|y_l|^b}{b}\right)\right).$$

By Young's inequality, given $1/a + 1/b = 1$, $|y_k|^a / a + |y_k|^b / b \geq y_k^2$, but equality must hold if $a = b = 2$. Similarly, $|y_l|^a / a + |y_l|^b / b \geq y_l^2$, but equality must hold if $a = b = 2$. Since $I_k / \pi_k \geq 0$ and $I_l / \pi_l \geq 0$, any choice $a \neq b$ can only yield $\widehat{\text{Var}}_{\text{Cab}}(\hat{t}) \geq \widehat{\text{Var}}_{C2}(\hat{t})$. Given all values of y_k and y_l, it is possible to derive an optimal vector of a_{kl} and b_{kl} values that varies over k, l, but such a derivation may not be of practical value.

4 Applications

Proposition 1 shows that the bias of the Horvitz-Thompson variance estimator under non-measurability is

$$A = \sum_{k \in U} \sum_{l \in \{U \setminus k : \pi_{kl} = 0\}} y_k y_l .$$

This expression, along with the fact that $A^* \geq A$, makes it evident that the degree of bias in $\widehat{\text{Var}}(\hat{t})$ and $\widehat{\text{Var}}_C(\hat{t})$ depends a great deal on the number of pairs with zero pairwise inclusion probabilities. For designs where this number is small, $\widehat{\text{Var}}(\hat{t})$ may provide a reasonable and conservative estimator for cases where y_k takes the same sign for all k, and $\widehat{\text{Var}}_C(\hat{t})$ may provide a reasonable and conservative estimator for cases where y_k may take different signs for some k. An example that arises frequently is stratified sampling where for a relatively small proportion of cases, we have small strata from which we draw only one unit.

For designs that result in many pairs having zero inclusion probabilities, $\widehat{\text{Var}}(\hat{t})$ and $\widehat{\text{Var}}_C(\hat{t})$ could be wildly over-conservative and other estimators may be preferred in terms of criteria such as mean square error. A prominent example is systematic sampling. Indeed, Särndal *et al.* (1992, page 76) propose that under systematic sampling, the Horvitz-Thompson variance estimator, $\widehat{\text{Var}}(\hat{t})$, can give a "non-sensical result." The expression for A makes it clear why this would be the case. Wolter (2007, Chapter 8) shows that simpler biased estimators, such as the with-replacement (Hansen-Hurwitz) variance estimator, can be reliable, if slightly conservative, in a broad range of data scenarios under equal probability and probability proportional to size (PPS) systematic sampling. Nonetheless, the with-replacement estimator fails to account adequately for sampling variance when outcome variance within systematic sample clusters is smaller than the between cluster variance. In such cases, $\widehat{\text{Var}}(\hat{t})$ would bound this variance in expectation when outcomes are all of the same sign, and $\widehat{\text{Var}}_C(\hat{t})$ would always bound this variance in expectation. Of course, it may still be the case that the bias is too large to be of much use, and so we would not suggest that $\widehat{\text{Var}}(\hat{t})$ and $\widehat{\text{Var}}_C(\hat{t})$ provides a full solution to the variance estimation problem for systematic sampling under high intra-cluster correlation.

Results from simulation studies are available in a supplement (at https://files.nyu.edu/cds2083/public/docs/smj_suppl.pdf). They illustrate how $\widehat{\text{Var}}(\hat{t})$ and $\widehat{\text{Var}}_C(\hat{t})$ perform relative

to commonly-used alternatives in applied scenarios. The simulations demonstrate situations when these estimators are preferable to the alternatives. For one-unit-per-stratum sampling, we show that these estimators are less biased than the "collapsed stratum" estimator in a range of scenarios. For PPS systematic sampling, these estimators perform favorably when the population exhibits substantial periodicity, a case when the commonly-used with-replacement estimator may be grossly negatively biased.

5 Conclusion

We have characterized precisely the bias of the Horvitz-Thompson variance estimator under non-measurability and used this characterization to develop a conservative bias correction. These estimators reflect the fundamental uncertainty inherent to non-measurable designs. Compared to available approximate methods, these estimators may sometimes perform better and sometimes worse from a practical perspective. But available approximate methods may be biased in ways that cannot always be evaluated in terms of either magnitude or sign. The estimators developed in this paper may therefore provide an informative measure of sampling variability with which analysts can agree without invoking additional assumptions or resorting to methods that carry the potential for negative bias. The bias term, A, has a simple form that suggests the possibility of refinements to the estimators developed here, something that we leave open for future research.

Acknowledgements

The authors wish to thank Joel Middleton, Allison Carnegie, an anonymous reviewer and the associate editor for helpful comments.

References

Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953). *Sample Survey Methods and Theory: Volume 1 Methods and Theory*. New York: John Wiley & Sons, Inc.

Horvitz, D., and Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 260, 663-685.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer.

Wolter, K. (2007). *Introduction to Variance Estimation, Second Edition*. New York: Springer.


ELECTRONIC PUBLICATIONS AVAILABLE AT
PUBLICATIONS ÉLECTRONIQUES DISPONIBLE À
www.statcan.gc.ca

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 28, No. 4, 2012

All inquires about submissions and subscriptions should be directed to jos@scb.se

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 29, No. 1, 2013

All inquires about submissions and subscriptions should be directed to jos@scb.se

The Canadian Journal of Statistics **La revue canadienne de statistique**

CONTENTS TABLE DES MATIÈRES

Volume 40, No. 4, December/décembre 2012

CONTENTS TABLE DES MATIÈRES

Volume 41, No. 1, March/mars 2013