Catalogue no. 12-001-X

# Survey Methodology


June 2011

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 at www.statcan.gc.ca, e-mail us at infostats@statcan.gc.ca, or telephone us, Monday to Friday from 8:30 a.m. to 4:30 p.m., at the following numbers:

**Statistics Canada's National Contact Centre**

Toll-free telephone (Canada and United States):

| | |
|---|---|
| Inquiries line | 1-800-263-1136 |
| National telecommunications device for the hearing impaired | 1-800-363-7629 |
| Fax line | 1-877-287-4369 |

Local or international calls:

| | |
|---|---|
| Inquiries line | 1-613-951-8116 |
| Fax line | 1-613-951-0581 |

**Depository Services Program**

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

## To access and order this product

This product, Catalogue no. 12-001-X, is available free in electronic format. To obtain a single issue, visit our website at 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<br>Finance<br>R.H. Coats Bldg., 6th Floor<br>150 Tunney's Pasture Driveway<br>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."

Statistics Canada

Business Survey Methods Division

# Survey Methodology

June 2011

Published by authority of the Minister responsible for Statistics Canada



June 2011

Catalogue no. 12-001-XIE
ISSN 1492-0921

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

Frequency: semi-annual

Ottawa

Cette publication est disponible en français sur demande (nº 12-001-X au catalogue).

---

**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 cooperation and goodwill.

# 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. 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).

**Subscription Rates**

The price of printed versions of *Survey Methodology* (Catalogue No. 12-001-XPB) is CDN $58 per year. The price excludes Canadian sales taxes. Additional shipping charges apply for delivery outside Canada: United States, CDN $12 ($6 × 2 issues); Other Countries, CDN $20 ($10 × 2 issues). A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, the Statistical Society of Canada and l'Association des statisticiennes et statisticiens du Québec. Electronic versions are available on Statistics Canada's web site: www.statcan.gc.ca.

# Survey Methodology

A Journal Published by Statistics Canada

Volume 37, Number 1, June 2011

## Contents

**Regular Papers**



**Short Notes**

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, June 2011
Vol. 37, No. 1, pp. 1-12
Statistics Canada, Catalogue No. 12-001-X



# Nonsampling errors in dual frame telephone surveys

**J. Michael Brick, Ismael Flores Cervantes, Sunghee Lee and Greg Norman** [1]

Abstract

Dual frame telephone surveys are becoming common in the U.S. because of the incompleteness of the landline frame as people transition to cell phones. This article examines nonsampling errors in dual frame telephone surveys. Even though nonsampling errors are ignored in much of the dual frame literature, we find that under some conditions substantial biases may arise in dual frame telephone surveys due to these errors. We specifically explore biases due to nonresponse and measurement error in these telephone surveys. To reduce the bias resulting from these errors, we propose dual frame sampling and weighting methods. The compositing factor for combining the estimates from the two frames is shown to play an important role in reducing nonresponse bias.

Key Words: Nonresponse bias; Measurement error; Calibration; Sample allocation; Composite.

## 1. Introduction

Dual frame telephone surveys that sample from both landline and cell phones have become important in the U.S. to reduce undercoverage bias due to the incompleteness of the landline frame. Blumberg and Luke (2009) show that the percentage of households without a landline telephone but with at least one cell phone has increased dramatically in the last few years, reaching 20 percent by the end of 2008. Other countries also report substantial increases in the percentages of people who have only a cell phone (*e.g.*, Kuusela, Callegaro and Vehovar 2008; Vicente and Reis 2009).

This paper uses data from the California Health Interview Survey (CHIS) and from 8 surveys conducted for the Pew Research Center for the People & the Press to examine the effects of nonsampling errors in dual frame telephone surveys. The CHIS 2007, a survey of California adults, was undertaken in late 2007. It combines a standard landline survey with a screening sample of cell phone numbers, where adults from the cell sample were interviewed only if they indicated that they did not have a landline number in the household. The Pew surveys are national surveys that interviewed an adult at all sampled residential telephone numbers from both landline and the cell samples. These surveys are described in more detail later. A number of important issues associated with the effect of nonsampling errors have been identified as a result of undertaking these dual frame telephone surveys – errors that have not been investigated fully in other studies.

In the next section we review sample design, weighting and variance estimation methods developed for dual frame surveys, and describe CHIS 2007 and Pew dual frame telephone surveys that are used throughout the paper. The third section discusses nonsampling error in dual frame telephone surveys, and the effects these errors may have on the bias of estimates. Nonresponse and measurement errors have special importance in dual frame surveys. The fourth section studies sampling and estimation methods that may be used to alleviate bias in dual frame telephone surveys, and gives conditions under which these sampling and estimation approaches may be most useful. In this section we propose three estimators to reduce the bias due to differential nonresponse within the overlap domain. The final section summarizes some of the findings for dual frame telephone surveys, and speculates on the applicability of these findings for other dual frame surveys.

## 2. Background

Most of the literature on dual frame surveys deals with the statistical theory related to efficiency in sample design and estimation. We summarize some of the key results in sampling, weighting and variance estimation, and then discuss the application of these methods to dual frame telephone surveys.

### 2.1 Sampling

The two sampling frames are denoted as $A$ and $B$, and we assume the samples from these frames, $S_A$ and $S_B$, are independent. The domain of units that are only in $A$ is $a$, the domain of units only in $B$ is $b$, and the intersection containing the overlap units is $ab$. In our application to telephone surveys, $A$ is the frame of landline numbers, $B$ is the frame of cell phone numbers, $a$ is the domain of households with only landline numbers, $b$ is the domain of households with only cell phone numbers, and $ab$ is the domain of households with both types of telephone service.

1. J. Michael Brick, Westat and Joint Program in Survey Methodology at the University of Maryland. E-mail: mikebrick@westat.com; Ismael Flores Cervantes, Westat; Greg Norman, Westat, 1600 Research Boulevard, Rockville Maryland, 20850 U.S.A.; Sunghee Lee, Institute for Social Research, University of Michigan, 426 Thompson St. Ann Arbor, MI 48104, U.S.A.

Many important features of dual frame surveys depend on how units that could fall into both sampling frames (*ab*) are handled.

A screening dual frame approach attempts to make $ab = \varnothing$ by removing any overlap units before sampling, after sampling but prior to data collection, during data collection, or after data collection. Lohr (2009) gives examples of dual frame surveys using each of these approaches.

Brick, Edwards and Lee (2007) and Fleeman (2007) describe screening in dual frame telephone surveys. While U.S. telephone numbers can be partitioned by whether they are cell or landline numbers, this frame does not identify whether those numbers correspond to households with only landlines (*a*), households with only cell phones (*b*), or households with both types of service (*ab*). In the surveys described by Brick, Edwards and Lee (2007) and Fleeman (2007), households sampled from the cell phone frame (*B*) were screened out during the data collection if they reported having a landline. The CHIS 2007 used this screening approach.

A second approach is called an overlap dual frame survey, and units in the overlap could be sampled from both frames. In this case, estimation methods must be employed to avoid biased estimates because the overlap units have multiple chances of selection. Steeh (2004), Brick, Brick, Dipko, Presser, Tucker and Yuan (2007), and Kennedy (2007) discuss dual frame telephone surveys with overlap. In these cases, all respondents are interviewed irrespective of the frame they are sampled from. The Pew surveys use the overlap approach.

### 2.2 Estimation

In a screening survey, producing weights for estimating totals and characteristics of the entire population is simple, at least in the absence of nonsampling errors. Since $ab = \varnothing$ and the sampling is independent, the units sampled from each frame are assigned weights that are the inverse of their selection probabilities from the frame from which they were selected. An overall estimate of the total is the sum of the weighted domain estimates, $\hat{y}_{scr} = \hat{y}_A + \hat{y}_b$, where $\hat{y}_A = \sum_{i \in S_A} d_i y_i$ and $\hat{y}_b = \sum_{i \in S_B} d_i \delta_i(b) y_i$, where $d_i$ is the inverse of the selection probability and $\delta_i(b) = 1$ if $i$ is in domain $b$ and 0 otherwise. Variance estimation is also straight-forward since the two frames are strata and variance estimation methods appropriate for stratified samples can be applied. For telephone surveys, the landline sample units are weighted and added to the weighted cell phone sampled units, after the sampled cell phone units that have landlines are given a weight of zero.

Screening during data collection, even in the absence of nonsampling errors, does have implications. For example, screened out households from *B* are not eligible for the interview, and this increases data collection costs and the variance of estimated totals (Kish 1965, Chapter 11). The units that are screened out should also be treated properly as sampled units in variance estimation.

Overlap surveys are more complex because units could be sampled from either of the frames. One estimation approach is to combine the two domain estimates, $\hat{y}_a$ and $\hat{y}_b$ with an average of the estimates of the overlap population from the separate frames. If $\hat{y}_{ab}^A$ and $\hat{y}_{ab}^B$ are the weighted estimates of the overlap domain from frame *A* and frame *B*, respectively, then an average or composite estimator is $\hat{y}_{ave} = \hat{y}_a + \hat{y}_b + \lambda \hat{y}_{ab}^A + (1 - \lambda)\hat{y}_{ab}^B$, with $0 \le \lambda \le 1$. Following Lohr (2009) we refer to these as average estimators. Assuming $\hat{y}_a$ and $\hat{y}_b$ are unbiased for domain *a* and domain *b*, and $\hat{y}_{ab}^A$ and $\hat{y}_{ab}^B$ are both unbiased for domain *ab*, then $\hat{y}_{ave}$ is an unbiased estimator of the total. Estimates of means and other quantities can be produced using weights, where the weights for units in *ab* that are sampled from *A* are multiplied by $\lambda$ and the weights for overlap units sampled from *B* are multiplied by $(1 - \lambda)$. The choice of the compositing factor, $\lambda$, has been investigated by many researchers and specific choices to reduce the variance of the estimates have been suggested by Hartley (1962, 1974) and Fuller and Burmeister (1972). All of average estimators require that the domain for all sampled units can be identified.

Variance estimation with the average estimator is relatively simple if $\lambda$ is a fixed and not dependent on the selected sample. In this case, $V(\hat{y}_{ave}) = V(\hat{y}_a + \lambda \hat{y}_{ab}^A) + V(\hat{y}_b + (1 - \lambda)\hat{y}_{ab}^B)$, and each of these variances can be computed using variance estimation methods appropriate for the separate samples. If $\lambda$ is sample dependent, as with the Hartley and Fuller and Burmeister estimators, then variance estimation is more complicated. The average estimators with a fixed $\lambda$ have been used in most dual frame telephone surveys with overlap. This approach is discussed below for the Pew surveys.

Other estimation approaches that have been considered for an overlap survey include the single frame estimator (Bankier 1986; Kalton and Anderson 1986; and Skinner 1991), and the pseudo-maximum-likelihood estimator (Skinner and Rao 1996; Lohr and Rao 2000; and Lohr and Rao 2006). Lohr (2009) reviews these estimators. Nearly all telephone surveys with overlap that we have seen use some versions of the average estimator, and it is the focus of this research.

### 2.3 Telephone survey applications

Data from CHIS 2007 are used to illustrate issues that arise in dual frame telephone survey that use a screening approach. The CHIS 2007 is a telephone survey of

California's population conducted by the UCLA Center for Health Policy Research in collaboration with the California Department of Public Health, the California Department of Health Care Services, and the Public Health Institute. Data collection for CHIS 2007 was carried out by Westat in late 2007 through early 2008.

In the CHIS 2007 landline sample, one adult was sampled and interviewed in each household. In the cell phone sample, persons living in households with landline phones were screened out; an adult was sampled and interviewed in the cell sample if they lived in a household classified as cell-only. All responding households, including those screened out from the cell phone frame, were asked questions about telephone status and usage. Nearly 49,000 adult interviews were completed from the landline sample, and 825 interviews were completed with cell-only adults. The landline sample response rate was 35.5% in the interview conducted with a household informant, and a 59.4% for the sampled adult. Respective response rates for the sample from the cell frame were 22.1% and 52.0%. Since CHIS 2007 used a screening approach, the reported response rate for the cell-only household informant interview is 30.5%. California Health Interview Survey (2009) discusses details of the study design, including differences between the overall cell phone response rate and the cell-only rate.

In the CHIS 2007, the estimates from the cell phone sample are calibrated to the cell-only adult population in California at the screening stage (prior to nonresponse weight adjustment for the sampled adult). There are some difficulties with obtaining reliable control totals for the calibration at the state level that are discussed later. The two samples from the two frames are independent samples and are treated as such, until the ultimate stage where the two are combined and calibrated to independent totals of the entire adult population of California. This last calibration stage does not include telephone status as a domain.

For dual frame telephone surveys with overlap, we use data aggregated from 8 surveys conducted for the Pew Research Center for the People & the Press in late 2008 through early 2009. (The data for the Pew surveys were provided by Scott Keeter of the Pew Research Center for the People & the Press). All of these are surveys of the entire U.S. adult population. The surveys interview one adult in each sampled household from both frames using nearly identical questionnaires. Over the 8 surveys, nearly 11,300 landline interviews and 3,800 cell phone interviews were completed. The response rates from the different surveys are very similar for the landline and the cell phone samples, with a median difference of one percentage point between the samples from the two frames. The response rates range across the 8 surveys and two frames from 17% to 24%.

In the Pew surveys, like most dual frame telephone surveys with overlap, a calibrated version of the average estimator is employed. Most surveys calibrate to both the telephone status domain counts (number of adults living in households with only cell phones, the number in household with only landlines, and households with both landlines and cell phones), and to demographic variables. The Pew studies are also calibrated to demographic totals including age, education, race/ethnicity, region, and population density of households with adults 18 years of age or older. In addition, they calibrate to totals of telephone status and, within the overlap domain to relative usage of landline and cell phones.

## 3. Nonsampling errors

Dual frame theory has been developed for ideal conditions – complete response and the absence of other nonsampling errors. Nonsampling errors affect the bias and precision of the estimates in any survey, but their effects in dual frame surveys may be qualitatively different from those in single frame surveys for three reasons. First, nonsampling error in dual frame surveys often makes it difficult to determine the probability of selection of the sampled unit. This occurs when domain membership is ascertained during data collection, and nonresponse and measurement errors make it difficult to determine if a sampled unit is in the overlap. Second, nonsampling error in dual frame surveys may be linked directly, sometimes causally, to the sampling frame especially when data collection approaches differ by frame. Third, sampling from more than one frame adds complexity and creates more opportunities for nonsampling errors to have differential effects.

### 3.1 Nonresponse effects

Brick, Dipko, Presser, Tucker and Yuan (2006) show that the over-representation of the number of adults in cell-only households that occurs in almost all dual frame telephone samples may be due to nonresponse error. They suggest that this over-representation might be the result of differential accessibility – adults who rarely use cell phones are less likely to answer their cell phone than those who use their cell phones regularly. They did not find the same type of usage-related differential response rates in the landline sample. Kennedy (2007) further explores this type of nonresponse bias by examining the effects on specific estimates.

To evaluate the differential representation, we compare the CHIS 2007 and Pew survey sample distributions by sampling frame and telephone usage to estimates from the National Health Interview Survey (NHIS). The NHIS is a

face-to-face survey sponsored by the National Center for Health Statistics with data collected by the U.S. Bureau of Census (the NHIS data were provided by S. Blumberg and J. Luke as a special tabulation). It is the only federal government survey that provides estimates of telephone status and usage (Blumberg and Luke 2009). We define usage for the dual users (those in households with both types of phone service) as cell-mainly and land-mainly, where cell-mainly are persons who live in households that receive all or almost all their calls on their cell phone and land-mainly are the dual users in households that do not receive all or almost all their calls on their cell phone.

To be more comparable to the CHIS figures, Table 1 restricts the NHIS estimates to those from the West region only (NHIS estimates for California are not available). California accounts for 52 percent of the adults in the West. The NHIS figures are population estimates from the first six months of 2008, which is roughly contemporaneous to the CHIS data collection period. The CHIS figures are the unweighted sample dispositions (the weighted dispositions are nearly identical). Even though CHIS used a screening approach, the telephone usage information was collected for every responding household in the cell phone sample. The table shows that the cell phone frame distribution over-represents the percent of adults in cell-only households and under-represents land-mainly adults when compared to the NHIS estimates. The landline respondents over-represent the land-only users and under-represent the cell-mainly dual users. The landline frame differences are more substantial than observed in a 2004 survey as reported in Brick *et al*. (2006).

Table 2 shows the same type of comparison of the NHIS national estimates from the second half of 2008 to the aggregated Pew survey unweighted outcomes (all the surveys were equal probability samples). Similar to the CHIS results, the cell frame distribution from the Pew surveys over-represents the percentage in the cell-only group and under-represents the land-mainly group, but the differences are less substantial than in CHIS. The Pew distribution from the landline sample mirrors the NHIS distribution closely, with a slight under-representation of the cell-mainly group.

**Table 1**
**Percentage distribution of adults from CHIS 2007 and NHIS, by telephone usage**

| Telephone usage | NHIS West adults in landline households | CHIS 2007 landline distribution | NHIS West adults in cell phone households | CHIS 2007 cell phone distribution |
|---|---|---|---|---|
| Landline-only | 23.5% | 34.2% | _ | _ |
| | (1.5%) | (0.2%) | | |
| Dual – land-mainly | 56.6% | 53.2% | 60.9% | 18.5% |
| | (1.7%) | (0.2%) | (1.7%) | (0.7%) |
| Dual – cell-mainly | 19.9% | 12.7% | 21.4% | 31.2% |
| | (1.4%) | (0.2%) | (1.4%) | (0.9%) |
| Cell-only | _ | _ | 17.7% | 50.3% |
| | | | (1.3%) | (0.9%) |
| Total | 100.0% | 100.0% | 100.0% | 100.0% |

Notes NHIS-West is the National Health Interview Survey, West Region, first 6 months of 2008, with percentages of all households with that type of service (thanks to S. Blumberg and J. Luke for this special tabulation). CHIS 2007 is the California Health Interview Survey, collected in 2007 and early 2008, with unweighted percentages from the landline and cell frames. In the cell phone sample, usage was obtained in the screening interview. Approximate standard errors given in ().

**Table 2**
**Percentage distribution of adults from Pew surveys and NHIS, by telephone usage**

| Telephone usage | NHIS adults in landline households | Pew surveys landline distribution | NHIS adults in cell phone households | Pew surveys cell phone distribution |
|---|---|---|---|---|
| Landline-only | 19.4% | 23.0% | _ | _ |
| | (0.7%) | (0.4%) | | |
| Dual – land-mainly | 58.8% | 62.7% | 58.8% | 42.3% |
| | (0.8%) | (0.5%) | (0.8%) | (0.8%) |
| Dual – cell-mainly | 19.3% | 14.4% | 18.5% | 24.0% |
| | (0.7%) | (0.3%) | (0.7%) | (0.7%) |
| Cell-only | _ | _ | 22.7% | 33.7% |
| | | | (0.7%) | (0.8%) |
| Total | 100.0% | 100.0% | 100.0% | 100.0% |

Notes NHIS is the National Health Interview Survey, second 6 months of 2008, with percentages of all households with that type of service. Pew surveys aggregates 8 surveys conducted for the Pew Research Center for the People & the Press from October 2008 through March 2009, with unweighted percentages from the landline and cell frames. (Thanks to S. Keeter for providing these data). Approximate standard errors given in ().

Both of these surveys exhibit response distributions by frame and usage that are consistent with the accessibility conjecture of Brick *et al.* (2006). This conjecture implies an ordering of those that are most accessible and likely to respond – ordering from the most likely to respond to the least likely to respond in the cell frame is cell-only, cell-mainly, and land-mainly. The special problem due to having two frames is that the ordering in the landline frame is different (land-only, land-mainly, cell-mainly), and the overlap units from the two frames could have very different response rates and biases.

To examine nonresponse bias for a dual frame survey with overlap, suppose both the landline and cell samples are poststratified to telephone status domain totals prior to forming an average overall estimate. The poststratified estimator is

$$\hat{y}_{ps} = \frac{N_a}{\hat{N}_a}\hat{y}_a + \frac{N_b}{\hat{N}_b}\hat{y}_b + \lambda g^A \hat{y}_{ab}^A + (1-\lambda) g^B \hat{y}_{ab}^B, \quad (1)$$

where the poststratification factor for the land-only sample is $N_a / \hat{N}_a$, for the cell-only sample it is $N_b / \hat{N}_b$, and the frame specific poststratification factors for the overlap are $g^A = N_{ab} / \hat{N}_{ab}^A$ and $g^B = N_{ab} / \hat{N}_{ab}^B$ for the landline and cell samples, respectively. The Horvitz-Thompson (HT) estimators of the number of units are $\hat{N}_a$ for the land-only domain, $\hat{N}_b$ for the cell-only domain, and $\hat{N}_{ab}^A$ and $\hat{N}_{ab}^B$ for the overlap domain from the two samples. Since we focus on the overlap, we write

$$\hat{y}_{ps,ab} = \lambda g^A \hat{y}_{ab}^A + (1-\lambda) g^B \hat{y}_{ab}^B. \quad (2)$$

This poststratified estimator differs from the approach suggested by Lohr and Rao (2000), who average and then poststratify rather than poststratify and then average. Both approaches are consistent and approximately unbiased when there are no nonsampling errors.

If we allow for differential response rates by telephone usage within the overlap such as those observed in dual frame telephone surveys, (2) is biased. Let $W$ be the proportion of the overlap that are land-mainly, and let $\overline{Y}_{ml}$ and $\overline{Y}_{mc}$ be the population means for a characteristic for land-mainly and cell-mainly dual users, respectively. The bias of $\hat{y}_{ps,ab}$ is

$$b(\hat{y}_{ps,ab}) \doteq WN_{ab}(\overline{Y}_{ml} - \overline{Y}_{mc}) (\lambda r_{l1} r_l^{-1} + (1-\lambda) r_{c1} r_c^{-1} - 1), \quad (3)$$

where $r_l$ is the dual user's response rate for the landline sample, $r_{l1}$ is the landline sample response rate of the land-mainly, $r_c$ is the dual user's response rate for the cell sample, and $r_{c1}$ is the cell phone sample response rate of the land-mainly.

To derive (3), we first define land-mainly and cell-mainly domain estimators from the landline sample as $\hat{y}_{ab}^A(ml) = \hat{N}_{ml}^A \overline{y}_{ab}^A(ml)$ and $\hat{y}_{ab}^A(mc) = \hat{N}_{mc}^A \overline{y}_{ab}^A(mc)$, and from the cell sample as $\hat{y}_{ab}^B(ml) = \hat{N}_{ml}^B \overline{y}_{ab}^B(ml)$ and $\hat{y}_{ab}^B(mc) = \hat{N}_{mc}^B \overline{y}_{ab}^B(mc)$. Now assume (a) $E\overline{y}_{ab}^A(ml) = E\overline{y}_{ab}^B(ml) = \overline{Y}_{ml}$ and $E\overline{y}_{ab}^A(mc) = E\overline{y}_{ab}^B(mc) = \overline{Y}_{mc}$; (b) covariances such as $\text{cov}(\hat{N}_{ml}^A / \hat{N}_{ab}^A, \overline{y}_{ab}^A(ml)) = 0$; and, (c) the expected domain totals are simple expressions such as $E\hat{N}_{ml}^A = r_{l1} N_{ml}$, $E\hat{N}_{ab}^A = r_l N_{ab}$, *etc*. Since $E(N_{ab} / \hat{N}_{ab}^A)\hat{y}_{ab}^A = N_{ab} E\{(\hat{N}_{ml}^A \overline{y}_{ab}^A(ml) + \hat{N}_{mc}^A \overline{y}_{ab}^A(mc)) / \hat{N}_{ab}^A\}$, we can write $E(N_{ab}/\hat{N}_{ab}^A)\hat{y}_{ab}^A \doteq r_{l1} r_l^{-1} N_{ml} \overline{Y}_{ml} + r_{l2} r_l^{-1} N_{mc} \overline{Y}_{mc} = N_{ab}(r_{l1} r_l^{-1} W(\overline{Y}_{ml} - \overline{Y}_{mc}) + \overline{Y}_{mc})$. A corresponding expression can be written for $Eg^B \hat{y}_{ab}^B$. Combining the two gives (3).

These expressions assume that $E\overline{y}_{ab}^B(ml) = \overline{Y}_{ml}$ and $E\overline{y}_{ab}^B(mc) = \overline{Y}_{mc}$. An alternative approach that does not require this assumption is to posit that there is response propensity associated with telephone usage. The bias in this case would be a function of the response propensities from each frame. We do not examine the response propensity approach here.

Expression (3) shows that when $0 < W < 1$, the bias of $\hat{y}_{ps,ab}$ is zero if (a) $\overline{Y}_{ml} = \overline{Y}_{mc}$; or (b) $\lambda r_{l1} r_l^{-1} + (1-\lambda) r_{c1} r_c^{-1} = 1$. Condition (a) is basically the well-known condition from single frame methodology. Condition (b) differs from single frame expressions because the bias depends on both the relative response rates and the compositing factor, $\lambda$. The exception is when $r_{l1} r_l^{-1} = r_{c1} r_c^{-1}$, or equivalently $r_{l1} r_{l2}^{-1} = r_{c1} r_{c2}^{-1}$, where $r_{l2}$ is the landline sample response rate of the cell-mainly and $r_{c2}$ is the cell sample response rate of the cell-mainly. In this form, this expression is comparable to the single frame bias expression that shows no bias exists when response rates are constant.

More generally, the value of $\lambda$ affects the bias of the estimate, not just its variance. The bias can be eliminated by choosing

$$\lambda_0 = \frac{r_l(r_c - r_{c1})}{r_c r_{l1} - r_l r_{c1}}. \quad (4)$$

Since the proportion of the total population covered by the landline frame is approximately equal to the proportion covered by the cell phone frame, most applications have used $\lambda = 0.50$ without considering its effect on bias.

We can now apply these expressions to evaluate the bias of dual frame telephone estimator for CHIS, assuming the bias is only from differential nonresponse in the overlap. Using the data in Table 1, $W = 0.74$ for the NHIS West region. We approximate $r_{l1} r_l^{-1}$ by the relative poststratification factor that is the ratio of the percentage of the CHIS landline sample classified as land-mainly to the percentage of the NHIS adults in landline households that are land-mainly; $r_{c1} r_c^{-1}$ is computed similarly for the cell phone

quantities. The quantities estimated from CHIS 2007 are given in Table 3, $r_{l1}\, r_l^{-1} \doteq 1.09$ for the landline sample, and $r_{c1}\, r_c^{-1} \doteq 0.50$ for the cell sample. As an example, suppose $\overline{Y}_{ml} = 0.3$ and $\overline{Y}_{mc} = 0.5$, then the bias of the estimated percentage based on (3) is approximately 3 percentage points (a relative bias of about 9%) if $\lambda = 0.5$. Using (4), the bias is zero when $\lambda \doteq 0.84$; the bias becomes negative for larger values of $\lambda$.

**Table 3**
**Within overlap, relative poststratification factors for CHIS 2007 and Pew surveys**

| Relative poststratification factors* | CHIS 2007 | Pew surveys |
|---|---|---|
| $r_{l1} r_l^{-1} \doteq g^A / g^A{}_{ml}$ | 1.09 | 1.07 |
| $r_{l2} r_l^{-1} \doteq g^A / g^A{}_{mc}$ | 0.50 | 0.84 |
| $r_{c1} r_c^{-1} \doteq g^B / g^B{}_{ml}$ | 0.74 | 0.78 |
| $r_{c2} r_c^{-1} \doteq g^B / g^B{}_{mc}$ | 2.42 | 1.51 |

* Poststratification adjustment factor for telephone usage domain within overlap divided by overlap poststratification factor.

The same computations can be done using the data from the Pew surveys, and the estimates are also shown in Table 3. The parameters differ substantially from those computed from CHIS. Since the Pew studies are national, the NHIS estimate is $W = 0.81$. The ratios of the Pew figures to the NHIS also have lower variability than those from the CHIS, with $r_{l1} r_l^{-1} \doteq 1.07$ and $r_{c1} r_c^{-1} \doteq 0.84$. As a result, the bias is only approximately 1 percentage points when $\lambda = 0.5$. The bias is zero when $\lambda \doteq 0.7$.

To evaluate the biases more completely, estimates of $\overline{Y}_{ml} - \overline{Y}_{mc}$ are needed for characteristics from a dual frame telephone survey rather than making arbitrary assumptions as done in the example above. Blumberg and Luke (2009) give estimates that suggest these differences may be as substantial as the differences between the cell-only and landline population that have been documented extensively elsewhere. However, the NHIS estimates are from a face-to-face survey, not a dual frame telephone survey.

Keeter, Dimock and Christian (2008) give estimated characteristics for dual telephone users by sampling frame, but not in sufficient detail to compute the biases. Keeter's estimates indicate the estimates of dual users from the cell frame might be closer to the NHIS overlap estimates than those from the landline frame. However, since the response rates within the overlap are more variable from the cell frame than from the landline frame, a screening design that aims to reduce bias should exclude dual users from the cell phone frame rather than the landline frame when the cell frame has more variable response rates by frame.

Because of the potential bias in the overlap design, Brick *et al*. (2006) suggest using a screening design that excludes adults in dual usage households if they were sampled from the cell frame. In a screening design, a bias still exists due to the differential nonresponse in the landline sample of dual users by telephone usage. Substituting $\lambda = 1$ into (2) and (3), the bias of $\hat{y}_{scr,ab} = g^A \hat{y}^A_{ab}$ is

$$b(\hat{y}_{scr,ab}) = WN_{ab}(\overline{Y}_{ml} - \overline{Y}_{mc})(r_{l1}\, r_l^{-1} - 1). \qquad (5)$$

The bias for this design and estimator is equivalent to single frame estimators, with the bias vanishing when either $\overline{Y}_{ml} = \overline{Y}_{mc}$ or the landline response rates are the same for the land-mainly and the cell-mainly. Notice that in this design, there is no compositing factor that can be used to control the bias.

The bias of the screener estimator for CHIS 2007 is about half that of the average estimator using $\lambda = 0.50$ (the screener bias is 1.3 percentage points compared to the post-stratified average estimator using $\lambda = 0.50$ with bias of -3.3 points). With the Pew parameters, the bias of the post-stratified average estimator and the screener estimator are nearly equal, with the bias of the screener slightly greater than the poststratified estimator (the screener bias is 1.1 percentage points compared to -0.7 points for the post-stratified overlap).

An issue mentioned earlier is that domain totals for poststratification, even for telephone status alone (land-only, cell-only, and dual domains), are not generally available for state or local area surveys. While small area estimates of the percentage of adults who are cell-only at the state level have been published (Blumberg, Luke, Davidson, Davern, Yu and Soderberg 2009), these do not give small area estimates for all three domains. The situation for telephone usage control totals is even more limited, with only national NHIS estimates published. Since the response rates in the cell frame typically vary by usage, some assumptions about the response rates in the cell sample may be useful to avoid substantial over-representation of cell-only and cell-mainly adults from the cell frame sample when using the overlap design.

### 3.2 Measurement error effects

In addition to nonresponse, some of the differences in the distributions shown in tables 1 and 2 could be due to measurement error. Before we discuss hypotheses related to measurement error, some of the key procedures in the surveys that could be related to measurement error are discussed. There are fundamental differences in the surveys, such as mode and topic. The NHIS is a face-to-face survey; the CHIS and Pew surveys are telephone surveys. Both NHIS and CHIS are health surveys, while the Pew surveys cover a broad range of topics.

The surveys also use different methods for collecting telephone status and usage. In the NHIS an adult family member is asked to answer questions about telephone status and usage for the entire family in a section of the interview about family characteristics. In the cell phone sample in CHIS 2007, the telephone status items are asked during the household screening, but the usage items are in the sampled adult interview. In the CHIS landline sample and the Pew surveys, the status and usage items are all in one of the last sections of the adult interview. This later placement is possible because no screening is involved.

The sampling of an adult is another procedure that may interact with the measurement process. In the CHIS 2007, an adult is sampled from all adults who share the same cell phone. In the Pew surveys, and most other cell phone surveys, the cell phone is considered a personal device, and the person answering the phone is interviewed. In dual use households, the CHIS and Pew methods may result in different samples of adults.

The greatest potential source of measurement error may be related to differences in the questionnaire items for telephone status and usage in the surveys. The items asked in each survey are given in the appendix. The approaches are quite varied. At least part of the difference in the studies is because the CHIS and Pew surveys are conducted by telephone and have prior information about telephone status.

The items used in all three surveys are derived from items used in a supplement to the Current Population Survey (CPS) in 2004. As discussed in Tucker, Brick and Meekins (2007), cognitive testing and behavioral coding for the supplement identified a number of concerns with the CPS items, especially the usage item. Their testing found that a lack of a specific reference period, not having a code for ''half the time,'' and difficulty in reporting for other members of the household made the usage item susceptible to measurement error. Tucker *et al*. (2007) also highlight the difficulty respondents had in reporting telephone status and usage for all household members in a single item. In addition, respondents had difficulty with understanding the meaning of "landline," "regular," a "working" cell phone, and the difference between using and answering a cell phone.

These issues could affect domain classification, and thus bias estimates. For example, a 23-year-old living with parents might report being cell-only, while the parents might report dual usage. The effects on the estimates of these types of measurement errors in the NHIS and telephone surveys are difficult to predict, but inconsistent reporting in telephone and face-to-face administrations is not unexpected.

Another possible measurement problem is the relationship between reporting telephone usage and the sampling frame from which respondents were selected. The hypothesized error arises if the respondent, when asked which device they use to receive most of their calls, is more likely to choose the device they are using to do the interview. We do not believe this hypothesis has been tested, but any device effect of this nature would be expected to be in the same direction as the nonresponse effect. A dual user should have a greater likelihood of reporting as cell-mainly if sampled from the cell frame; they should be more likely to report as land-mainly if sampled from the landline. Thus, the bias discussed earlier in the context of nonresponse could be arising due to the combined effect of nonresponse and device effect. Without being able to identify the magnitude of these sources of the bias, methods for reducing bias are unclear.

## 4. Design and estimation approaches with nonsampling errors

Because of the additional issues at play in dual frame surveys, sampling and estimation methods should be designed to account for the most important sources of error rather than focusing solely on sampling error. In this section we address sample design and estimation choices for dual frame telephone surveys within this larger error structure setting.

### 4.1 Sample design approaches

A key design decision for a dual frame telephone survey is whether to use a screening or full overlap sample design. We begin by exploring the optimal allocation of the sample for overlap and screening designs appropriate for dual frame telephone surveys when simple random samples are selected independently from the two frames and $N_a > 0$, $N_b > 0$, and $N_{ab} > 0$. We assume throughout that the sample sizes are large enough to ignore the finite population correction factors.

We use a linear expected cost function $E(C) = c_A(n_A + n_B c_B c_A^{-1})$, where $c_A$ is the cost of a landline interview, $c_B$ is the cost of a cell phone interview, and $n_A$ and $n_B$ are the number sampled from frames $A$ and $B$, respectively. Assuming a constant element variance, $\sigma^2$, the variance of the overlap estimator is $v_{ov}^2 = \sigma^2(N_A(N_a + \lambda^2 N_{ab})n_A^{-1} + N_B(N_b + (1-\lambda)^2 N_{ab})n_B^{-1})$. The allocation that minimizes the variance with this cost function can be found by standard Lagrangian methods, and is

$$n_{o,A} = E(C)\tau^{-1}\sqrt{c_A^{-1}N_A(N_a + \lambda^2 N_{ab})}$$

$$n_{o,B} = E(C)\tau^{-1}\sqrt{c_B^{-1}N_B(N_b + (1-\lambda)^2 N_{ab})}, \quad (6)$$

where

$$\tau = \sqrt{c_A N_A(N_a + \lambda^2 N_{ab})} + \sqrt{c_B N_B(N_b + (1-\lambda)^2 N_{ab})}.$$

For a screening design, a linear cost function appropriate for dual frame telephone surveys is $E(C) = c_A n_A + n_b c_b$, where $c_b = c_B + N_B N_b^{-1} c_s$, $n_b$ is the sampled number of cell-only, and $c_s$ is the cost of screening. The variance of the screening estimator is $v_{sc}^2 = \sigma^2(N_A^2 n_A^{-1} + N_B N_b n_B^{-1})$. The optimal allocation is just the stratified allocation given by $n_{s,A} = E(C)N_A(c_A N_A + \sqrt{c_A c_b} N_b)^{-1}$ and

$$n_B = \frac{E(C)N_B}{\sqrt{c_A c_b} N_A + c_b N_b},$$

yielding

$$n_b = \frac{E(C)N_b}{\sqrt{c_A c_b} N_A + c_b N_b}$$

cell-only interviews.

With no nonsampling error and a fixed expected cost, the variance for the optimally allocated overlap design is smaller than the variance for the optimally allocated screener design when the cost of screening is large enough so that $\sqrt{c_b} > N_b^{-1}(\tau - N_A\sqrt{c_A})$. When bias is included, the screening design may have smaller mean square error than the overlap design even when this condition holds. In the analysis below, we consider bias but do not account for all the effects of nonsampling error. For example, differential response affects the yield by the sampling frame from which the units are selected thus affecting the allocation and variance of the estimate.

We compare the mean square errors of the screening and overlap designs under the CHIS 2007 parameters given previously. The mean square error is the sum of the variance and the bias squared. The variance is for the overall estimate, but the bias arises only from the overlap under our assumptions. The cost parameters for interviewing and screening cell phones are still not very well-known, but we use $(c_A = 1, c_B = 3, c_s = 2)$ based on information given by Keeter *et al.* (2008) and Edwards, Brick and Grant (2008). The other parameters needed for the comparison are the distribution of the population by telephone status domain, and we approximate national values from the 2008 NHIS national estimates ($N_a = 0.2N$, $N_b = 0.2N$, and $N_{ab} = 0.6N$). In this situation, the variance based on an optimally allocated overlap design with $\lambda = 0.5$ is slightly smaller than the variance for the optimal screening design (the ratio of the variances is 0.976). The variances of the two designs are approximately the same when the cost parameters are such that the screening from frame *B* is slightly less expensive ($c_A = 1, c_B = 3, c_s = 1.85$).

The screening approach has smaller mean square error than the overlap design under these conditions because the screening approach reduces the bias of the estimates from -3.3 percentage points to 1.3 points. Even a relatively small bias dominates the mean square error comparison between the two designs, assuming the bias with the screening approach is half the bias under the overlap design. This is the case because the variances of the overlap and screening designs are so similar. If we instead use the parameters from the Pew surveys, then the mean square error for the overlap design is smaller because its bias is lower than the bias of the screener design.

The allocation to the frames with the overlap approach given by (6) assuming only sampling error is determined by the population parameters, the cost parameters, and the compositing factor. While this is not the optimal allocation when differential response rates are admitted, it is still useful to consider this situation since it is likely to be encountered frequently in practice. In this situation, the bias of $\hat{y}_{ps,ab}$ due to differential nonresponse can be eliminated by choosing $\lambda$ to satisfy (4). Based on the CHIS parameters, the value that eliminates this bias is $\lambda \doteq 0.84$. If we continue with the cost and population assumptions as above, but set $\lambda = 0.84$, then the optimal allocation given by (6) would select about 75% of the sample from the landline frame. This contrasts with the allocation with $\lambda = 0.5$, in which only 63% is from the landline frame. The choice of the compositing factor is critical. When $\lambda = 0.84$ is used in conjunction with the optimal allocation for the CHIS parameters, the estimator is unbiased and has a variance that is about 5 percent less than the estimator from the optimal screener design.

### 4.2 Estimation approaches

An approach suggested by Brick *et al.* (2006) is to use a full overlap design with an average estimator for the overlap that is poststratified to telephone usage domain totals, as is done in the Pew surveys. This estimator is unbiased and consistent if the estimates within the domains are unbiased and the domain sample sizes are sufficiently large.

The auxiliary data needed for this poststratification for the entire U.S. are now published regularly from the NHIS. As mentioned above, there are some concerns about using these data as control totals that deserve further study. The control totals needed for this estimator are the number of land-only adults, the number of cell-only adults, and the number of adults who are land-mainly and the number who are cell-mainly ($N_{ml}$ and $N_{mc}$, respectively). This partitions the dual users into its two components.

An alternative estimator of the overlap total using the same auxiliary data is

$$\begin{aligned}\hat{y}_{sep} = {} & \frac{N_a}{\hat{N}_a}\hat{y}_a + \frac{N_b}{\hat{N}_b}\hat{y}_b + \lambda_1 g_{ml}^A \hat{y}_{ab}^A(ml) \\ & + (1-\lambda_1) g_{ml}^B \hat{y}_{ab}^B(ml) \\ & + \lambda_2 g_{mc}^A \hat{y}_{ab}^A(mc) + (1-\lambda_2) g_{mc}^B \hat{y}_{ab}^B(mc), \end{aligned} \quad (7)$$

where the detailed poststratification factors are $g^A_{ml} = N_{ml}/\hat{N}^A_{ml}$, $g^A_{mc} = N_{mc}/\hat{N}^A_{mc}$, $g^B_{ml} = N_{ml}/\hat{N}^B_{ml}$, $g^B_{mc} = N_{mc}/\hat{N}^B_{mc}$, and $0 \le \lambda_1 \le 1$; $0 \le \lambda_2 \le 1$. This estimator, like the others considered thus far, is unbiased and consistent in the absence of nonsampling errors. Like (1), the estimates from each frame are poststratified before being averaged. The primary difference between (1) and (7) is that the dual users in (7) are partitioned and poststratified by usage; it also introduces different compositing factors within the overlap.

The estimator $\hat{y}_{sep}$ may be useful when (1) is biased and usage control totals are available for poststratification. If the expected means within the usage domains are approximately equal ($E\bar{y}^A_{ab}(ml) = E\bar{y}^B_{ab}(ml) = \bar{Y}_{ml}$ and $E\bar{y}^A_{ab}(mc) = E\bar{y}^B_{ab}(mc) = \bar{Y}_{mc}$), then (7) is unbiased for any choice of $0 \le \lambda_1 \le 1$ and $0 \le \lambda_2 \le 1$. Since bias is not affected by the choice, different compositing factors may be used to reduce the variance of the estimates as is traditionally suggested in the dual frame literature. Table 3 shows that the proportion of respondents in the detailed usage domains varies considerably by the sampling frame, and this might make different compositing factors worthwhile.

Because telephone usage control totals often are not available, we explored modifying (2) to use different compositing factors similar to those used in the overlap for (7). In this case, the goal would be to reduce bias rather than variance. A modified estimator of the overlap total is

$$\hat{y}_{\mathrm{mod},ab} = \lambda_1 g^A \hat{y}^A_{ab}(ml) + (1-\lambda_1) g^B \hat{y}^B_{ab}(ml) + \lambda_2 g^A \hat{y}^A_{ab}(mc) + (1-\lambda_2) g^B \hat{y}^B_{ab}(mc). \quad (8)$$

However, this estimator may not be useful for reducing bias. Earlier, we showed that the bias of $\hat{y}_{ps,ab}$ vanishes when $\lambda_0 = r_l (r_c - r_{c1})(r_c r_{l1} - r_l r_{c1})^{-1}$. The choice of $\lambda_1 = \lambda_2 = \lambda_0$ in (8) eliminates the bias for both land-mainly and cell-mainly estimates, so that different compositing factors are not useful for bias reduction. The bias of the modified estimator is

$$b(\hat{y}_{\mathrm{mod},ab}) = WN_{ab}(\bar{Y}_{ml}(\lambda_1 r_{l1} r_l^{-1} + (1-\lambda_1) r_{c1} r_c^{-1} - 1) - \bar{Y}_{mc}(\lambda_2 r_{l1} r_l^{-1} + (1-\lambda_2) r_{c1} r_c^{-1} - 1)), \quad (9)$$

where we make assumptions similar to those used earlier to approximate the bias of $\hat{y}_{ps,ab}$.

Another reason for studying an overlap estimator like (8) is because it is appropriate with sample designs that screen out land-mainly adults from the cell frame. This approach has been considered because the number of cell frame respondents that are classified as land-mainly may be small, and the assumption that $E\hat{y}^B_{ab}(ml) = \bar{Y}_{ml}$ may not hold and biases might result.

Setting $\lambda_1 = 1$, (8) reduces to

$$\hat{y}_{\mathrm{mod}\,\lambda=1,ab} = g^A \hat{y}^A_{ab}(ml) + \lambda_2 g^A \hat{y}^A_{ab}(mc) + (1-\lambda_2) g^B \hat{y}^B_{ab}(mc). \quad (10)$$

In this design, the landline sample alone is used to estimate both the land-only and the land-mainly totals. Both frames are used to estimate totals for the cell-mainly. If we assume $E\bar{y}^A_{ab}(ml) = \bar{Y}_{ml}$ and $E\bar{y}^A_{ab}(mc) = E\bar{y}^B_{ab}(mc) = \bar{Y}_{mc}$, then we no longer need $E\bar{y}^B_{ab}(ml) = \bar{Y}_{ml}$ for (10) to be unbiased. As before, setting $\lambda_2 = r_l (r_c - r_{c1})(r_c r_{l1} - r_l r_{c1})^{-1}$ eliminates the bias in the cell-mainly estimate.

## 5. Discussion

This exploration of nonresponse and measurement errors in dual frame telephone surveys suggests the effects of these errors may be very important. It leads us to believe that research on nonsampling errors to reduce biases may be more important than research that leads to incremental reductions in sampling error.

The research also reveals shortcomings in our knowledge about nonsampling errors in these surveys. The direction and magnitude of the effects of measurement error are especially unclear. The inconsistencies in some of the findings for the CHIS 2007 and Pew surveys may well be due to measurement errors associated with the different approaches to data collection in these surveys, or to interactions due to the procedures. A thorough investigation of the error sources in dual frame telephone surveys is essential to improve the quality of dual frame telephone surveys, and we believe experiments to assess the effects of measurement error would be especially beneficial.

We did find that the CHIS 2007 and Pew surveys consistently over-represented cell-only and cell-mainly users in samples from the cell phone frame, and the surveys had a slight over-representation of the land-only and land-mainly from the landline frame. However, the degree of over-representation of the domains differed by survey. In the CHIS, the over-representation could have led to substantial biases in the estimates if an overlap survey and a simple average estimator were used. The CHIS used a screening approach to reduce this potential bias, and this appears to have been largely successful. In the Pew surveys, the representation was less differential by frame and the potential for bias was smaller. In these conditions, the overlap approach may have smaller mean square error than a screening approach.

Due to the potential for bias in dual frame telephone surveys with response patterns like the CHIS 2007, we examined sampling and estimation methods that could be implemented to deal with these biases. We found that screening approaches may be competitive or even preferable in dual frame telephone surveys when the bias due

to differential nonresponse or measurement error is large. If the bias is not negligible, this finding even holds with small sample sizes. However, these results depend on the choice of the compositing factor and the current practice of choosing $\lambda = 0.5$ should be reconsidered. An alternative is to choose the compositing factor to eliminate the bias of the average estimator. In many cases, this approach not only eliminates the bias, but also may be more efficient.

We examined three estimators that deal with the bias due to differential nonresponse within the overlap domain. The first is $\hat{y}_{ps}$, which uses telephone status as domain control totals. This estimator eliminates the bias due to differential nonresponse when $\lambda_0$ is used as the compositing estimator. This compositing factor indirectly uses information on the land-mainly and cell-mainly domain totals in computing response rates by domain and frame. A second estimator, $\hat{y}_{sep}$, eliminates this source of bias more directly by post-stratifying to telephone status and usage control totals. This estimator also permits the use of different compositing factors within the overlap domain to reduce the variance of the estimates. The third estimator that might be used to reduce bias is $\hat{y}_{\text{mod}}$, but this estimator is more pertinent for a sample design that interviews the cell-only and the cell-mainly respondents from the cell frame, along with all respondents from the landline sample. This modified screening design and estimator might be especially attractive if there is concern that the mean of the land-mainly respondents from the cell frame sample is subject to nonresponse bias. All of these estimators could also be raked to additional demographic control totals after combining the two samples.

Given our current state of knowledge, we believe there are important advantages with the full overlap design and $\hat{y}_{ps}$ with $\lambda_0$ chosen based on other similar surveys. It is worth observing that even though the CHIS and Pew surveys had very different response patterns, choosing a value of $\lambda_0 = 0.75$ would have reduced the bias substantially for both surveys. An advantage of this estimator over $\hat{y}_{sep}$ in general is that $\hat{y}_{ps}$ is not poststratified to usage domain totals. We suspect that usage domain totals estimated from a face-to-face survey (NHIS) may be subject to substantially different errors than the estimates from telephone surveys. These differences could result in telephone survey estimates that are biased and have underestimated variances. For state and local surveys where even telephone status totals are not well-known, control totals for usage domains are likely to be highly suspect.

A screening design with $\hat{y}_{scr}$ as the estimator has the advantage that it only requires control totals for the entire population and for the cell-only component, such as those estimated from the NHIS. A disadvantage is that, unlike the overlap estimators, there is no compositing parameter that can be used to reduce the bias directly. The more elaborate screening design that interviews cell-only and cell-mainly from the cell frame and uses $\hat{y}_{\text{mod}}$ has merit, but there have been no studies that examine the conditions which would favor this estimator.

A more complete analysis of the effects of nonsampling error would include other factors such as the effect of the differential response rates by frame. For example, we noted that samples from the cell phone frame yield more cell-only households than would be expected. These differential response rates can be addressed in allocating the sample, but we have not done so here. Our exploration of this shows that it results in larger allocations to the landline frame, increases the value of the compositing factor, and makes the screening designs more efficient relative to the overlap designs. The screening design and estimator are still subject to the bias noted above.

While this research concentrated on nonsampling errors in dual frame telephone surveys, we suspect that similar issues exist in many other dual frame surveys, but that these issues may not be recognized. Lohr (2009) mentions nonsampling errors in general dual frame surveys and suggests comparing estimates of the overlap from each frame as a simple diagnostic test. We believe this is an excellent way to begin an investigation of problems associated with the overlap.

As we noted earlier, the handling of the overlap is a major concern in dual frame surveys because nonsampling error may be associated with the sampling frame. Our investigation shows that nonresponse and measurement errors are tied to the sampling frame in dual frame telephone surveys. It is very likely that dual frame telephone surveys that use different modes might experience analogous effects. For example, consider a dual frame household survey designed to survey members of a rare population. Suppose it uses an incomplete membership list with telephone numbers for the rare group as frame *A*, and an area probability sample of households as frame *B*. Different response rates by sampling frame within the overlap might be expected, and these might be related to characteristics of the respondents leading to biases. Even within the overlap, there may be differences such as those related to how long the person has been a member of the organization used to create frame *A* and this might be related to characteristics such as age. This type of situation might parallel some of the within overlap domain issues identified in telephone surveys. Differential measurement errors related to the modes are also possible.

Given the potential for bias in a dual frame survey, one of the important findings of our research is that the compositing factor, $\lambda$, influences the bias as well as having an effect on the variance. While the choice of $\lambda$ typically has only a slight effect on the variance if $\lambda$ is in the vicinity of

the optimal value, the bias may be more sensitive to this choice. Thus, in dual frame surveys understanding how the choice of $\lambda$ affects the bias and the mean square error of the estimates is an important consideration. The other sampling and estimation methods discussed in this paper may also be applicable to other dual frame surveys. The usefulness of these methods depends upon understanding the nature of the nonsampling errors as well as the availability of auxiliary data that could be used in calibration.

## Acknowledgements

We would like to thank Scott Keeter, Stephen Blumberg and Julian Luke for providing data for this paper. We would also like to thank many people for helpful comments on earlier drafts including Sherm Edwards, Ralph DiGaetano, David Grant, David Hubble, Paul Lavrakas, Graham Kalton, Scott Keeter, and Courtney Kennedy.

## Appendix

## Telephone usage items

### National Health Interview Survey

N1. *Is there at least one telephone inside your home that is currently working and is not a cellular phone*?

N2. *Does anyone in your family have a working cellular telephone*?

N3. *How many working cellular telephones do people in your family have*?

[If both N1 and N2 are 'yes' ask N4]

N4. *Of all the telephone calls that your family receives, are …*

*All or almost all calls received on cell phones*?

*Some received on cell phones and some on regular phones*?

*Very few or none received on cell phones*?

### California Health Interview Survey – Cell phone

CC1. *Is this cell phone your only phone or do you also have a regular telephone at home*?

[If the phone is a cell phone and they have a regular phone then ask CC2]

CC2. *Of all the telephone calls that you receive, are …*

*All or almost all calls received on cell phones*

*Some received on cell phones and some on regular phones, or*

*Very few or none on cell phones*?

[If respondent replies about half, record it]

### California Health Interview Survey – Landline

CL1. *Do you have a working cell phone*?

[If yes or they share a cell phone ask CL2]

CL2. *Of all the telephone calls that you receive, are …*

*All or almost all calls received on cell phones*

*Some received on cell phones and some on regular phones, or*

*Very few or none on cell phones*?

[If respondent replies about half, record it]

### Pew Research Center for the People & The Press – Cell phone

PC1. *Now thinking about your telephone use… Is there at least one telephone INSIDE your home that is currently working and is not a cell phone*?

[If yes ask PC2]

PC2. *Of all the telephone calls that you receive, do you get*?

[Rotate options−keeping SOME in the middle]

*All or almost all calls on a cell phone*

*Some on a cell phone and some on a regular home phone*

*All or almost all calls on a regular home phone*

### Pew Research Center for the People & The Press – Landline

PL1. *Now thinking about your telephone use… Do you have a working cell phone*?

[If yes ask PL2]

PL2. *Of all the telephone calls that you receive, do you get*?

[Rotate options−keeping SOME in the middle]

*All or almost all calls on a cell phone*

*Some on a cell phone and some on a regular home phone*

*All or almost all calls on a regular home phone*

## References

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

Blumberg, S.J., and Luke, J.V. (2009). Wireless Substitution: Early Release of Estimates from the National Health Interview Survey, July-December 2008. National Center for Health Statistics. Available at http://www.cdc.gov/nchs/nhis.htm.

Blumberg, S.J., Luke, J.V., Davidson, G., Davern, M.E., Yu, T. and Soderberg, K. (2009). Wireless substitution: State-level estimates from the National Health Interview Survey, January–December 2007. Hyattsville, MD: National Center for Health Statistics. *National Health Statistics Reports*, 14.

Brick, J.M., Brick, P.D., Dipko, S., Presser, S., Tucker, C. and Yuan, Y. (2007). Cell phone survey feasibility in the U.S.: Sampling and calling cell numbers versus landline numbers. *Public Opinion Quarterly*, 71, 29-33.

Brick, J.M., Dipko, S., Presser, S., Tucker, C. and Yuan, Y. (2006). Nonresponse bias in a dual frame sample of cell and landline numbers. *Public Opinion Quarterly*, 70, 780-793.

Brick, J.M., Edwards, W.S. and Lee, S. (2007). Sampling telephone numbers and adults, interview length, and weighting in the California Health Interview Survey cell phone pilot study. *Public Opinion Quarterly*, 71, 793-813.

California Health Interview Survey (2009). CHIS 2007 Methodology Series: Report 4 – Response Rates. Los Angeles, CA: UCLA Center for Health Policy Research. Available at www.chis.ucla.edu/pdf/CHIS2007_method4.pdf.

Edwards, W.S., Brick, J.M. and Grant, D. (2008). Relative Costs of a Multi-frame, Multi-mode Enhancement to an RDD Survey. Presented at the Annual Conference of the American Association for Public Opinion Research, New Orleans, LA.

Fleeman, A. (2007). Survey Research Using Cell Phone Sample: Important Operational and Methodological Considerations. Presented at the Annual Conference of the American Association for Public Opinion Research, Anaheim, CA.

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

Hartley, H.O. (1962). Multiple Frame Surveys. *ASA Proceedings of the Social Statistics Section*, 203-206.

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

Kalton, G., and Anderson, D.W. (1986). Sampling Rare Populations. *Journal of the Royal Statistical Society*, A 149, 65-82.

Kish, L. (1965). *Survey Sampling*. New York : John Wiley & Sons, Inc.

Kennedy, C. (2007). Evaluating the effects of screening for telephone service in dual frame RDD Surveys. *Public Opinion Quarterly*, 71, 750-771.

Keeter, S., Dimock, M. and Christian, L. (2008). Calling Cell Phones in '08 Pre-election Polls. News Release from The Pew Research Center for the People & the Press. Available at www.pewresearch.org/pubs/1061/cell-phones-election-polling.

Kuusela, V., Callegaro, M. and Vehovar, V. (2008) The influence of mobile telephones on telephone surveys. In *Advances in Telephone Survey Methodology*, (Eds., J.M. Lepkowski, C. Tucker, J.M. Brick, E.D. de Leeuw, L. Japec, P.J. Lavrakas, M.W. Link and R.L. Sangster), New York: John Wiley & Sons, Inc., Chapter 4, 87-112.

Lohr, S. (2009). Multiple frame surveys. In *Handbook of Statistics*: *Sample Surveys_Design*, *Methods and Applications*, (Ed., D. Pfeffermann). Elsevier, Amsterdam, Chapter 4, Vol. 29A.

Lohr, S., and Rao, J.N.K. (2000). Inference in dual frame surveys. *Journal of the American Statistical Association*, 95, 271-280.

Lohr, S., and Rao, J.N.K. (2006). Estimation in multiple-frame surveys. *Journal of the American Statistical Association*, 101, 1019-1030.

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

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

Steeh, C. (2004). A New Era for Telephone Surveys. Presented at the Annual Conference of the American Association for Public Opinion Research, Phoenix, AZ.

Tucker, C., Brick, J.M. and Meekins, B. (2007). Household telephone service and usage patterns in the U.S. in 2004: Implications for telephone samples. *Public Opinion Quarterly*, 71, 3-22.

Vicente, P., and Reis, E. (2009). The mobile-only population in Portugal and its impact in a dual frame telephone survey. *Survey Research Methods*, 3, 105-111.

Survey Methodology, June 2011
Vol. 37, No. 1, pp. 13-24
Statistics Canada, Catalogue No. 12-001-X



# Maximum likelihood estimation for contingency tables and logistic regression with incorrectly linked data

**James O. Chipperfield, Glenys R. Bishop and Paul Campbell** [1]

Abstract

Data linkage is the act of bringing together records that are believed to belong to the same unit (*e.g.*, person or business) from two or more files. It is a very common way to enhance dimensions such as time and breadth or depth of detail. Data linkage is often not an error-free process and can lead to linking a pair of records that do not belong to the same unit. There is an explosion of record linkage applications, yet there has been little work on assuring the quality of analyses using such linked files. Naively treating such a linked file as if it were linked without errors will, in general, lead to biased estimates. This paper develops a maximum likelihood estimator for contingency tables and logistic regression with incorrectly linked records. The estimation technique is simple and is implemented using the well-known EM algorithm. A well known method of linking records in the present context is probabilistic data linking. The paper demonstrates the effectiveness of the proposed estimators in an empirical study which uses probabilistic data linkage.

Key Words: Data linkage; Probabilistic linkage; Maximum likelihood; Contingency tables; Logistic regression.

## 1. Introduction

Data linking, also referred to as data linkage or record linkage, is the act of bringing together records that are believed to belong to the same unit (*e.g.*, a person or business), from two or more files. Data linkage is an appropriate technique when data sets must be joined to enhance dimensions such as time and breadth or depth of detail. Ideally, the linkage will be perfect, meaning only records belonging to the same unit are linked and all such links are made. However, in many situations this does not happen, especially when linking records using fields that may have incorrect values, missing values or values that are legitimately different for a given unit.

Probabilistic linking is often used when the files contain a set of common variables or fields that constitute partial identifying information, but which do not constitute a unique unit identifier. In probabilistic linking (Fellegi and Sunter 1969) all possible links are given a score based on the probability that the records belong to the same unit. This score is calculated by comparing the values of linking variables that are common to both files. A link is then declared if the link score is higher than some cut-off. An optimisation algorithm may be used to ensure that each record on one file is linked to no more than one record on the other file. Probabilistic methods for linking files are now well established (see Herzog, Scheuren and Winkler 2007, Winkler 2001 and Winkler 2005) and there is a range of computer packages available to implement them.

This is a consequence of the continued importance of linkage in a variety of fields, particularly relating to health and social policy. Recent examples of probabilistic data linkage from the Australian Bureau of Statistics (ABS) include linking records from the 2006 Australian Census of Population and Housing to a number of data sets including Australian death registrations (Australian Bureau of Statistics 2008), the 2006 Census Dress Rehearsal (Solon and Bishop 2009), and the Australian Migrants Settlements Database (Wright, Bishop and Ayre 2009). In the health arena within Australia, probabilistic linkage methods are used by the Western Australian Data Linkage Unit (Holman, Bass, Rouse and Hobbs 1999) and by the New South Wales Centre for Heath Record Linkage. Internationally, probabilistic methods are used by Statistics Canada (Fair 2004), USBC (see Winkler 2001), the U.S. National Center for Health Statistics (National Center for Health Statistics 2009) and by the Switzerland Statistical agency as part of their Longitudinal Study of People Living in Switzerland.

Data linking offers opportunities for new statistical output and analysis. Naively treating a probabilistically-linked file as if it was perfectly linked will, in general, lead to biased estimates. Lahiri and Larsen (2005) and Scheuren and Winkler (1993) proposed methods to calculate unbiased estimates of coefficients for a linear regression model under probabilistic record linkage. More recently, Chambers, Chipperfield, Davis and Kovačević (2009) and Chambers (2008) extended this work to a wide set of models using generalised estimating equations and, in the case of linking two files, allowing one file to be a subset of the other file.

This paper develops a maximum likelihood (ML) approach for analysis of probabilistically-linked records. The estimation technique is simple and is implemented using the well-known EM algorithm. The approach involves replacing the statistics, which would be observed from perfectly linked

1. James O. Chipperfield, Australian Bureau of Statistics. E-mail: james.chipperfield@abs.gov.au; Glenys R. Bishop, The Australian National University; Paul Campbell, Australian Bureau of Statistics.

data, with their expectation conditional on the linked data. Assuming this expectation is correctly specified, this approach overcomes the following two limitations of the previous work.

First, the previous methods assume only one linkage pass is made, whereas, probabilistic linkage usually involves multiple passes. In the latter case, records not linked in the first pass are eligible to be linked in the second pass, and only records not linked in the first two passes are eligible to be linked in the third pass, and so on. Each pass is designed to link records with a particular common set of characteristics. For example, the first pass may be designed to link records belonging to individuals who have not changed address between the reference dates of the two files. The second pass may be designed to accommodate changes of address. An example of such an approach is given in Table 1 in section 5.

Second, the previous methods assume that either the two files contain records from exactly the same units or the set of units on one file is a subset of those on the other file. The approach proposed can be used when one of the files to be linked is not necessarily a subset of the other file. This situation occurs frequently in practice and occurred in all the ABS examples mentioned above. It is also worth mentioning that the files to be linked do not need to be related via a sampling mechanism, such as the smaller file being a random sub-sample of individuals from the larger file. Removing this restriction means that the two files may be administrative data sets.

Consider linking two files denoted by X and Y. File Y contains the variable $y$ on the population of individuals $U_y$ comprising $n_y$ records. File X contains a vector of variables, $\mathbf{x}$, on the population of individuals $U_x$ comprising $n_x$ records. The target of inference is with respect to the population of $n_{xy}$ individuals, denoted by $U_{xy} = U_x \cap U_y$, who are common to File X and File Y. Files X and Y also contain a vector of fields, denoted by $\mathbf{z}$, which are used to link the files using a probabilistic linkage algorithm. Of course, since we are considering probabilistic linkage here, the variable $\mathbf{z}$ does not constitute a unique unit identifier.

Linking Files X and Y allows the joint distribution of $\mathbf{x}$ and $y$ to be analysed. There are two sources of error that may affect analysis of the joint distribution using the linked file. These errors are referred to as *incorrect links* and *unlinked records*.

A link is correct when the pair of linked records belong to the same individual. A link is incorrect when a pair of linked records do not belong to the same individual. Incorrect links can artificially increase or decrease the correlation between $\mathbf{x}$ and $y$. An example of the latter is random linkage, where records on File X are randomly linked to records on File Y.

The $i^{\text{th}}$ record on File X is defined as an *unlinked record*, if $i \in U_{xy}$ *and* record $i$ was not linked to a record on File Y. Or in other words, an unlinked record is a record on File X that could be correctly linked but was not linked at all (throughout this paper we use the convention of defining unlinked records in terms of File X, though the definition could equally be in terms of records on File Y). It may not always be possible to link a particular record on File X with much confidence that the link is correct. This situation may arise if a record is missing fields that are useful in establishing the correct link. More generally, unlinked records may occur when some sub-populations are relatively difficult to link. For example, fields such as marital status, qualification, field of study, and highest level of schooling would generally not be as powerful when linking children as when linking mature adults. In this situation, the data linker must decide whether or not to link such records. We define the set of linked records by $U_l$ of size $n^*$ so that $n^* \leq n_x$ and $n^* \leq n_y$.

The problem of analysis with unlinked records has clear parallels with the problem of unit non-response. Both lead to only a subset of legitimate records being available for analysis. The non-response mechanism in survey sampling is, in reality, a function of an unknown set of variables. Here however, we have the slight advantage in knowing that the probability of a record remaining unlinked can only be a function of $\mathbf{z}$. The problem of non-response is often addressed by weighting or by some conditioning argument. This paper considers both approaches to address the issue of unlinked records.

There is a natural trade-off between the number of unlinked records and incorrect links (and consequently the bias that they introduce). Consider the case where File X is a subsample of File Y so that $U_{xy} = U_x$. Linking all records on File X will result, by definition, in no unlinked records but will result in the number of incorrect links being maximised. If instead we decide to only form links which we are very confident are correct, the number of incorrect links will decrease but the number of unlinked records will increase. In practice, finding the optimal balance between the biases due to unlinked records and incorrect links depends upon the analysis to be undertaken, the linkage methodology, and their interaction. For an in-depth practical discussion of this issue see Bishop (2009).

It is worthwhile mentioning that the problem of making inference in the presence of incorrect record linkage is similar to the problem of making inference in the presence of misclassification of the outcome variable, which is a form of measurement error (see Fuller 1987). In the latter case, identifying assumptions separate the misclassification mechanism from the model mechanism and are required since no error-free measurement is typically available. For example,

Hausman, Abrevaya and Scott-Morton (1998) considers misclassification in the outcome variable of a logistic regression model. Their identifying assumption is that the value of the, possibly misclassified, outcome variable is a particular function of the model's explanatory variables. Our proposed method does not require the strong identifying assumptions of measurement error problems essentially because error-free measurement is available from a clerical sample which identifies correct links. The assumptions we make in this paper are outlined in section 3.

Section 2 summarises the ML approach to contingency table and regression analysis under perfect linkage. Section 3 considers the ML approach in the presence of incorrect links. Section 4 considers the ML approach in the presence of both incorrect links and unlinked records. Section 5 demonstrates the effectiveness of many of the proposed estimators in an empirical study. Section 6 summarises the findings.

## 2. Perfect linkage

By way of introducing notation, this section discusses the case where the linkage is perfect. The estimating approach in this section is standard since, clearly, no special adjustment for incorrect linkage is required. Section 2.1 discusses estimating cell probabilities in a contingency table and section 2.2 discusses estimating regression coefficients in a logistic regression.

### 2.1 Contingency tables

For notation, it is convenient when considering contingency table analysis to transform $\mathbf{x}_i$ to a single categorical variable $x$ so that $x = 1, 2, ..., g, ..., G$. Define $y$ to be a categorical variable on file Y, where $y = 1, ..., c, ..., C$.

Consider the following factorisation of the distribution of x and $y$

$$p(y,\ x) = p_1(y \mid x;\ \mathbf{\Pi})\, p_2(x),$$

where $\mathbf{\Pi} = (\boldsymbol{\pi}_1', ..., \boldsymbol{\pi}_g', ..., \boldsymbol{\pi}_G')'$, $\boldsymbol{\pi}_g = (\pi_{1|g}, ..., \pi_{c|g}, ..., \pi_{C|g})'$, $\pi_{c|g}$ is the probability that $y = c$ given $x = g$. We assume that for every value of $x$ there are $C$ possible values of $y$ which implies that the dimension of $\mathbf{\Pi}$ is $CG$.

We now consider maximum likelihood estimation of the parameter $\mathbf{\Pi}$, characterising $p_1$, under perfect linkage. Perfect linkage means that all records on file X are correctly linked to their corresponding record on file Y (*i.e.*, there are no incorrect links and no unlinked records). Under perfect linkage, $n_{xy} = n_x$ and the set of linked records is denoted by $\mathbf{d} = \{(y_i,\ x_i)\colon i = 1, ..., n_{xy}\}$. Under perfect linkage, the score function for $\boldsymbol{\pi}_x = (\pi_{1|x}, ..., \pi_{c|x}, ..., \pi_{C|x})'$ characterised by the multinomial distribution, is

$$\text{Score}(\boldsymbol{\pi}_x;\ \mathbf{d}) =$$

$$(\text{Score}(\pi_{1|x};\ \mathbf{d}), ..., \text{Score}(\pi_{c|x};\ \mathbf{d}), ..., \text{Score}(\pi_{C-1|x};\ \mathbf{d}))' \quad (1)$$

where

$$\text{Score}(\pi_{c|x};\ d) = \Sigma_i (w_{ic|x}\pi_{ic|x}^{-1} - w_{iC|x}\pi_{iC|x}^{-1})$$

$$= n_{c|x}\pi_{c|x}^{-1} - n_{C|x}\pi_{C|x}^{-1},$$

for $c = 1, ..., C-1$, where $n_{c|x} = \Sigma_i w_{ic|x}$, $w_{ic|x} = 1$ if $y_i = c$ and $x_i = x$ and $w_{ic|x} = 0$ otherwise, and the category corresponding to $y = C$ is the arbitrarily chosen reference category. Solving $\text{Score}(\boldsymbol{\pi}_x;\ \mathbf{d}) = \mathbf{0}_{C-1}$ for $\boldsymbol{\pi}_x$, where $\mathbf{0}_{C-1}$ is a $C-1$ column vector of zeros, gives the maximum likelihood (ML) estimator

$$\hat{\pi}_{c|x} = n_{c|x}/n_x, \quad (2)$$

where

$$n_x = \Sigma_c \Sigma_i w_{ic|x}$$

and

$$\hat{\pi}_{C|x} = 1 - \sum_{c=1}^{C-1} \hat{\pi}_{c|x}.$$

### 2.2 Logistic regression

Consider the logistic regression model

$$E(y_i) = \upsilon_i \quad (3)$$

$$\upsilon_i = 1 / [1 + \exp(\boldsymbol{\beta}'\mathbf{x}_i)]. \quad (4)$$

For (4) the $K$ elements of $\mathbf{x}_i$ are dichotomous variables and $y_i$ is now a dichotomous variable available from File Y. If we define $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_i, ..., \mathbf{x}_{n_{xy}})'$, $\mathbf{y} = (y_1, ...y_i, ..., y_{n_{xy}})'$ and $\boldsymbol{\upsilon} = (\upsilon_1, ...\upsilon_i, ..., \upsilon_{n_{xy}})'$, the score matrix for $\boldsymbol{\beta}$ based on perfectly linked data, $\mathbf{d}$, is

$$\text{Score}(\boldsymbol{\beta};\ \mathbf{d}) = \mathbf{x}'(\mathbf{y} - \boldsymbol{\upsilon}). \quad (5)$$

Solving $\text{Score}(\boldsymbol{\beta};\ \mathbf{d}) = \mathbf{0}_K$ for $\boldsymbol{\beta}$ gives the ML estimate $\hat{\boldsymbol{\beta}}$, which can be found by applying the well-known Newton-Raphson method.

## 3. Analysis with incorrect links

This section considers the situation where the linked file contains incorrect links but does not contain unlinked records. This occurs when all the records on File X are linked to a record on File Y (so $n_x \leq n_y$). Define the linked file of records by $\mathbf{d}^* = \{\mathbf{d}_i^* = (y_i^*, \mathbf{x}_i)\colon i = 1, ..., n_x\}$, where $y_i^*$ is the value of $y$ that is *linked* to record $i$ on file X. To clarify, $y_i$ is the true value of $y$ for record $i$ on file X, so that $y_i^* = y_i$ if record $i$ is correctly linked.

The estimator given by (2), together with the assumption that $y_i^* = y_i$ for $i = 1, ..., n_x$, is naive since it treats the probabilistically linked file as if it were perfectly linked. In general the naive estimator will be biased. This section derives ML estimators which account for the fact that the data have been linked probabilistically or linked imperfectly in some way.

It is common practice to select a subsample of the linked file, denoted by $s_c$, which is then reviewed clerically. The clerical review classifies a link, $\mathbf{d}_i$, as either correct or incorrect. Let $\delta_i = 1$ if record *i* on File X is correctly linked and $\delta_i = 0$ otherwise.

Designing the clerical subsample is an important problem, especially since clerical review is often a costly exercise. Possible uses of a clerical sample include estimating the proportion of correctly linked and unlinked records, to assist in deciding which records should be linked and which should remain unlinked, to ensure correct inference using $\mathbf{d}^*$ (*i.e.*, the purpose of this paper), and to identify improvements to the way in which records are linked (in the ABS applications mentioned above, clerical samples were designed to ensure that each link had at least a specific probability of being correct). For the purpose of making correct inference using $\mathbf{d}^*$ selecting the clerical sample by simple random sampling is a reasonable approach. A more efficient clerical subsample could possibly be devised but there is no obvious way to do so. This is because the parameters that we need to estimate to implement the ML method described in this paper depend upon the specific analysis (*e.g.*, choice of $y$ and $\mathbf{x}$). Designing a clerical sample for all possible analyses would be difficult.

We factorise the joint distribution $p(y_i, \mathbf{x}_i, \delta_i)$ by

$$p(y_i \mid \mathbf{x}_i; \boldsymbol{\theta})\, p(\mathbf{x}_i)\, p(\delta_i \mid \mathbf{x}_i), \tag{6}$$

where $\boldsymbol{\theta} = \boldsymbol{\beta}$ in the regression case, $\boldsymbol{\theta} = \boldsymbol{\Pi}$ in the contingency table case. Factorisation (6) means that the links are incorrect at random (IAR) or, in other words, that the distributions $y_i \mid \mathbf{x}_i$ and $\delta_i \mid \mathbf{x}_i$ are independent. Under this assumption it is only necessary to maximise the likelihood associated with the factor $p(y_i \mid \mathbf{x}_i; \boldsymbol{\theta})$. Throughout this section we assume (6). It is important to point out that (6), and the development that follows, makes no assumption requiring File X to be a subset of File Y (*e.g.*, when units on File X are a subsample of the units on File Y) or that the linkage process involves a single pass. We also assume that the correctness of linkage, $\delta_i$, is independent from record to record.

As mentioned in the introduction, each linked record is assigned a score based on the probability that the records belong to the same unit. Denote the score by $r_i$. A referee suggested using $r_i$ to more accurately parameterise the distribution of $\delta_i$. Technically this suggestion would involve replacing $p(\delta_i \mid \mathbf{x}_i)$ with $p(\delta_i \mid \mathbf{x}_i, r_i)$ in (6) and would likely reduce the variability of the ML estimators discussed in section 3. This would be a useful avenue of further research.

### 3.1 Contingency tables

Define $w_{ic|x}^* = 1$ if $y_i^* = c$ and $x_i = x$, and $w_{ic|x}^* = 0$ otherwise. The expectation of $w_{ic|x}$ given $\mathbf{d}_i^*$ is

$$\begin{aligned} E_{d|d^*}(w_{ic|x} \mid x_i = x, y_i^* = y^*) = & \\ w_{ic|x}^* p_{xy^*} + (1 - p_{xy^*})\pi_{c|x} \quad & \text{if } i \notin s_c \\ = w_{ic|x}^* \quad & \text{if } i \in s_c \text{ and } \delta_i = 1 \\ = \pi_{c|x} \quad & \text{if } i \in s_c \text{ and } \delta_i = 0 \end{aligned}$$

and $p_{xy^*}$ is the probability that the $i^{\text{th}}$ link is correct given $x_i = x$ and $y_i^* = y^*$. The ML estimator of $\pi_{c|x}$ using the probabilistically linked data, $\mathbf{d}_i^*$, is then

$$\tilde{\pi}_{c|x} = \tilde{n}_{c|x} \left( \sum\nolimits_c \tilde{n}_{c|x} \right)^{-1} \tag{7}$$

where

$$\tilde{n}_{c|x} = \Sigma_i \tilde{w}_{ic|x}, \tag{8}$$

$$\begin{aligned} \tilde{w}_{ic|x} = w_{ic|x}^* \hat{p}_{xy^*} + (1 - \hat{p}_{xy^*})\tilde{\pi}_{c|x} \quad & \text{if } i \notin s_c \\ = w_{ic|x}^* \quad & \text{if } i \in s_c \\ = \tilde{\pi}_{c|x} \quad & \text{if } i \in s_c \text{ and } \delta_i = 0 \end{aligned} \tag{9}$$

and

$$\hat{p}_{xy^*} = \left( \sum\nolimits_{i \in s_c} w_{ic|x}^* \delta_i \right) \left( \sum\nolimits_{i \in s_c} w_{ic|x}^* \right)^{-1}. \tag{10}$$

The estimation procedure involves iterating between (7), (8) and (9) until convergence. Specifically the algorithm is:

1. Calculate $\hat{p}_{xy^*}$ from (10).
2. Initialise $\tilde{\pi}_{c|x}^{(0)}$ and then calculate $\tilde{w}_{c|x}^{(0)}$ from (9) and then $\tilde{n}_{c|x}^{(0)}$ from (8).
3. Calculate $\tilde{\pi}_{c|x}^{(t)}$ from (7) using $\tilde{n}_{c|x}^{(t-1)}$.
4. Calculate $\tilde{w}_{c|x}^{(t)}$ from (9) using $\tilde{\pi}_{c|x}^{(t)}$ and then calculate $\tilde{n}_{c|x}^{(t)}$ from (8) using $\tilde{w}_{c|x}^{(t)}$.
5. Iterate between 3 and 4 until convergence.

The initialised value $\tilde{\pi}_{c|x}^{(0)}$ could be set to the naive estimate of $\pi_{c|x}$, which was described in section 3 above. However, our experience was that the choice of initial value was not important.

### 3.2 Logistic regression

Below we describe two ML methods (Methods 1 and 2) for estimating $\boldsymbol{\beta}$ using the probabilistically linked data, $\mathbf{d}^*$.

Both methods give unbiased estimates under the IAR assumption. The difference between the methods is the level of aggregation at which the probabilities of correct linkage are estimated. Method 1 requires these probabilities at a fine level of aggregation, which may mean its estimates are more variable than those of Method 2.

### 3.2.1 Method 1

The expectation of $y$ conditional on the linked data is

$$E_{\mathrm{d}|\mathrm{d}^*}(y_i \mid \mathbf{x}_i = \mathbf{x},\, y_i^* = y^*) =$$
$$y_i^* p_{\mathbf{x}y^*} + (1 - p_{\mathbf{x}y^*})\upsilon_i \quad \text{if } i \notin s_c$$
$$= y_i^* \quad \text{if } i \in s_c \text{ and } \delta_i = 1$$
$$= \upsilon_i \quad \text{if } i \in s_c \text{ and } \delta_i = 0$$

and $p_{xy^*}$ is the probability that the $i^{\text{th}}$ link is correct given $x = x_i$ and $y_i^* = y^*$.

The ML estimator is then obtained by iterating between finding the solution, denoted by $\tilde{\boldsymbol{\beta}}$, for $\boldsymbol{\beta}$ in (5) with $y_i$ replaced by $\tilde{y}_i$, where

$$\begin{aligned} \tilde{y}_i &= y_i^* \hat{p}_{\mathbf{x}y^*} + (1 - \hat{p}_{\mathbf{x}y^*})\tilde{\upsilon}_i \quad \text{if } i \notin s_c \\ &= y_i^* \quad \text{if } i \in s_c \text{ and } \delta_i = 1 \\ &= \tilde{\upsilon}_i \quad \text{if } i \in s_c \text{ and } \delta_i = 0, \end{aligned} \tag{11}$$

$\tilde{\upsilon}_i$ has the same form as $\upsilon_i$ except that $\boldsymbol{\beta}$ is replaced with $\tilde{\boldsymbol{\beta}}$ and $\hat{p}_{\mathbf{x}y^*}$ is the estimated proportion of correct links in the clerical sample for each combination of $\mathbf{x}$ and $y^*$.

### 3.2.2 Method 2

Let $\mathbf{x}'\mathbf{y}$ in (5) have $k^{\text{th}}$ element

$$r_k = \mathbf{x}_k'\mathbf{y} = \sum_i^n y_i x_{ik} = \sum_i^n r_{ik},$$

where $r_{ik} = y_i x_{ik}$. The expectation of $r_{ik}$ conditional on $\mathbf{d}^*$ is

$$\begin{aligned} &E_{\mathrm{d}|\mathrm{d}^*}(r_{ik} \mid \mathbf{x}_i = \mathbf{x},\, y_i^* = y_i) = \\ &\quad [y_i^* p_{ky^*} + (1 - p_{ky^*})\upsilon_i]x_{ik} \quad \text{if } i \notin s_c \\ &\quad = y_i^* x_{ik} \quad \text{if } i \in s_c \text{ and } \delta_i = 1 \\ &\quad = \upsilon_i x_{ik} \quad \text{if } i \in s_c \text{ and } \delta_i = 0 \end{aligned} \tag{12}$$

and $p_{ky^*}$ is the probability that a link with $x_{ik} = 1$ is correct given $y_i^* = y^*$. The ML estimator is then obtained by iterating between finding the solution, denoted by $\tilde{\tilde{\boldsymbol{\beta}}}$, for $\boldsymbol{\beta}$ in (5) with $r_{ik}$ replaced by $\tilde{r}_{ik}$, where

$$\begin{aligned} \tilde{r}_{ik} &= [y_i^* \hat{p}_{ky^*} + (1 - \hat{p}_{ky^*})\tilde{\tilde{\upsilon}}_i]x_{ik} \quad \text{if } i \notin s_c \\ &= y_i^* x_{ik} \quad \text{if } i \in s_c \text{ and } \delta_i = 1 \\ &= \tilde{\tilde{\upsilon}}_i x_{ik} \quad \text{if } i \in s_c \text{ and } \delta_i = 0, \end{aligned} \tag{13}$$

$\tilde{\tilde{\upsilon}}_i$ has the same form as $\upsilon_i$ except that $\boldsymbol{\beta}$ is replaced with $\tilde{\tilde{\boldsymbol{\beta}}}$ and $\hat{p}_{ky^*}$ is the estimated proportion of correct links in the clerical sample for each combination of $\mathbf{x}$ and $y^*$. Namely, if $y_i^* = 1$,

$$p_{ky^*} = \left(\sum_{i \in s_c}^n y_i^* x_{ik}\delta_i\right)\left(\sum_{i \in s_c}^n y_i^* x_{ik}\right)^{-1}$$

and if $y^* = 0$,

$$p_{ky^*} = \left(\sum_{i \in s_c}^n (1 - y_i^*) x_{ik}\delta_i\right)\left(\sum_{i \in s_c}^n (1 - y_i^*) x_{ik}\right)^{-1}.$$

This approach requires only $2K$ probabilities to be calculated from the clerical sample and, on this basis, may be preferable to the approach in section 3.2.1 which requires more probabilities to be calculated.

### 3.3 Estimating the variance using the bootstrap

In this section we describe how to calculate the variance of the ML estimates of section 3. Denote the parameter of interest by $\boldsymbol{\theta}$, introduced earlier, and its ML estimate by $\tilde{\boldsymbol{\theta}}$. The Bootstrap (Rubin and Little 2003) estimate of the variance of $\tilde{\boldsymbol{\theta}}$, denoted by $\hat{v}_{\text{boot}}(\tilde{\boldsymbol{\theta}})$, is obtained by

1. Taking a replicate sample of size $n_x$ from the linked file, $\mathbf{d}^*$, by simple random sampling with replacement. Denote the $r^{\text{th}}$ replicate sample by $\mathbf{d}^*(r)$. The $r^{\text{th}}$ replicate clerical sample is $s_c(r) = s_c \cap \mathbf{d}^*(r)$.
2. Calculating $\tilde{\boldsymbol{\theta}}(r)$ which has the same form as $\tilde{\boldsymbol{\theta}}$ except that $\mathbf{d}^*(r)$ is used instead of $\mathbf{d}^*$ and $s_c(r)$ is used instead of $s_c$.
3. Repeating steps 1 and 2 $R$ times, where $R$ is the number of replicates.
4. Calculating

$$\hat{v}_{\text{boot}}(\tilde{\boldsymbol{\theta}}) = \frac{1}{R}\sum\nolimits_{b=1}^{R}(\tilde{\boldsymbol{\theta}}(b) - \tilde{\boldsymbol{\theta}})(\tilde{\boldsymbol{\theta}}(b) - \tilde{\boldsymbol{\theta}})'.$$

## 4. Analysis with incorrect links and unlinked records

This section discusses two ways of analysing linked data in the presence of incorrect links and unlinked records. As mentioned in the introduction, the problem of analysis when there are unlinked records has clear parallels with the problem of unit non-response. Unlinked records may result in some characteristics on the linked file being over- or under-represented, thus leading to biased analysis. As discussed in more detail below, we use the fact that the mechanism giving rise to unlinked records can only be a function of $\mathbf{z}$.

This section considers two methods of making inference in the presence of incorrect links and unlinked records, where linked records are indexed by $i = 1, ..., n^*$. (Remember that the $i^{th}$ record on File X is an *unlinked record* if $i \in U_{xy}$ *and* record $i$ was not linked to any record on File Y.) The methods involve independently modelling the processes that determine which records are incorrectly linked and which are unlinked (see section 5 for an illustration). These models require a subsample, denoted by $s_{xc}$, of all records on File X to be subjected to clerical review. Records in the subsample will be either linked to records on File Y or not linked. Linked records in the subsample must be identified as either correctly or incorrectly linked by the clerical review process. A subsample record which is not linked must be identified as either *unlinked*, or *otherwise*. *Unlinked* means the corresponding record was found on File Y but not linked to it, whereas *otherwise* indicates the corresponding record was not found on File Y and therefore assumed not to exist. The latter identification is potentially much more difficult and time-consuming than the former because it assumes some other error-free process is available for checking whether links, which were not made, are in fact correct. Unlinked records, by their nature, have limited information that can be used to identify the correct link, even during clerical review. Such a process may not exist, in which case adjusting for unlinked records would seem to be impossible. However, such a process may involve a clerical review of names appearing on the two files to be linked. For example, a clerical reviewer may realise that the names *John O. Smith* and *Joh O. Smith* on two different records may in fact be the same name (with an "n" missing in the latter case, perhaps due to errors in scanning), whereas the automated linking process may treat the two names as completely different. The clerical reviewer may then decide that the above two records correspond to the same individual and so therefore should be linked. (Bishop (2009) and Wright (2009) discuss the benefits of clerical review).

The first method involves conditioning analysis on a variable $\zeta_i = \zeta_i(\mathbf{z}_i)$. The variable $\zeta$ is defined so that inference, in the presence of unlinked records, is unbiased conditional on $\zeta$. The term $\zeta$ is introduced since, in many cases, it would be impractical or unnecessary to condition on all the information in $\mathbf{z}$. It is possible to give $\zeta_i$ a non-missing value even when $\mathbf{z}_i$ contains missing values. The exact form of the function $\zeta(\mathbf{z})$ would need to be justified after analysis of the subsample, $s_{xc}$. For example, if persons under 20 years of age are under-represented in the linked file, $\zeta$ would indicate whether a person is under 20 years of age. One approach to analysis is to include $\zeta$ as a covariate in the regression model. The method in section 3 would then apply directly. However, analysts may like to integrate over $\zeta$ so that it does not appear in the logistic model or contingency table. Section 4.2 discusses how to do this for contingency tables. Section 4.3 discusses a pseudo-likelihood approach which assigns weights to the linked records that attempt to account for any under- or over-representation of certain subpopulations in the linked data. Again, the choice of weight would need to be justified after analysis of the subsample, $s_{xc}$, which identifies unlinked records. This is discussed further in the context of the empirical study.

### 4.1 Can we ignore unlinked records?

Define the variable $\gamma_i = 1$ if record $i$ on File X is unlinked and $\gamma_i = 0$ otherwise. Also let $\zeta_i$ be a variable so that $\zeta_i = 1, 2, ..., h, ...H$, where $H$ is the number of categories for $\zeta$. We can ignore the fact that there are unlinked records if we are prepared to assume that, conditional on $\mathbf{x}_i$, the distributions of $y_i$, $\gamma_i$ and $\delta_i$ are independent. Technically this assumption leads to the factorisation,

$$p(y_i, \mathbf{x}_i, \delta_i, \gamma_i, \zeta_i) \propto$$
$$p(y_i \mid \mathbf{x}_i; \boldsymbol{\theta})\, p(\delta_i \mid \mathbf{x}_i)\, p(\gamma_i \mid \mathbf{x}_i)\, p(\zeta_i)$$

where again $\boldsymbol{\theta} = \boldsymbol{\beta}$ or $\boldsymbol{\Pi}$. It is worthwhile checking whether this assumption is valid from the clerical subsample. If the assumption is reasonable, then there is no need to apply the methods in section 4.2 and 4.3 and the methods in section 3 will suffice.

We may not be prepared to make the assumption mentioned above. We may however be prepared to assume, conditional on $\mathbf{x}$ and $\zeta$, the distributions of $y_i$, $\gamma_i$ and $\delta_i$ are independent. In this case, we say unlinked records are not ignorable. Technically this assumption leads to the factorisation,

$$p(y_i, \mathbf{x}_i, \delta_i, \gamma_i, \zeta_i) \propto$$
$$p(y_i \mid \mathbf{x}_i, \zeta_i; \boldsymbol{\Lambda})\, p(\delta_i \mid \mathbf{x}_i; \boldsymbol{\tau})\, p(\gamma_i \mid \mathbf{x}_i, \zeta_i)\, p(\zeta_i)$$

where $\boldsymbol{\Lambda}$ is the parameter for the distribution of $y_i \mid \mathbf{x}_i, \zeta_i$. If we are interested in $p(y_i \mid \mathbf{x}_i; \boldsymbol{\theta})$ but not $p(y_i \mid \mathbf{x}_i, \zeta_i; \boldsymbol{\Lambda})$, one approach is to integrate out (*i.e.*, average over) $\zeta_i$ from the latter.

### 4.2 Conditional Maximum Likelihood (CML) for contingency tables

First. parameterise the joint distribution of $y_i$, $x_i$ and $\zeta_i$ by the multinomial distribution with parameter, $\boldsymbol{\Lambda}$. Define $\boldsymbol{\Lambda} = (\boldsymbol{\Pi}_1', ... \boldsymbol{\Pi}_h', ..., \boldsymbol{\Pi}_H')'$, where $\boldsymbol{\Pi}_h = (\boldsymbol{\pi}_{1h}', ..., \boldsymbol{\pi}_{gh}', ..., \boldsymbol{\pi}_{Gh}')'$, $\pi_{gh} = (\pi_{1|gh}, ..., \pi_{c|gh}, ..., \pi_{C|gh})'$ and $\pi_{c|gh}$ is the probability that $y_i = c$, $x_i = g$ and $\zeta_i = h$. The ML estimator of $\boldsymbol{\Pi} = (\pi_{c|x})$ from section 2.1 when linkage errors are not ignorable is $\tilde{\tilde{\boldsymbol{\Pi}}} = (\tilde{\tilde{\pi}}_{c|x})$, where

$$\tilde{\tilde{\pi}}_{c|x} = \sum_{h=1}^{H} \tilde{\pi}_{c|xh} \hat{\pi}_{h|x}, \qquad (14)$$

where

$$\tilde{\pi}_{c|xh} = \tilde{\mathrm{n}}_{c|xh}\left(\sum_c \tilde{\mathrm{n}}_{c|xh}\right)^{-1}, \quad (15)$$

$\tilde{\mathrm{n}}_{c|xh} = \Sigma_{i\in U_l}\tilde{w}_{ic|xh}$, $\Sigma_{i\in U_l}$ is the sum over the $n^*$ linked records and $\hat{\pi}_{h|x}$ for $h = 1, ..., H$ is the standard estimate of the marginal distribution of $\zeta$ given $x$ on File X. Further, if $i \notin s_c$

$$\tilde{w}_{ic|xh} = w^*_{ic|xh}\hat{p}_{xy^*h} + (1 - \hat{p}_{xy^*h})\,\tilde{\tilde{\pi}}_{c|xh}, \quad (16)$$

$\hat{p}_{xy^*h}$ is the probability that the $i^{\text{th}}$ link is correct given $x_i = x$, $\zeta_{\mathrm{i}} = h$ and $y^*_i = y^*$, $w^*_{ic|xh} = 1$ if $x_i = x$, $\zeta_{\mathrm{i}} = h$ and $y^*_i = y^*$, and $w^*_{ic|xh} = 0$ otherwise. If $i \in s_c$, then $\tilde{w}_{ic|xh} = w^*_{ic|xh} = w_{ic|xh}$ if the link is determined to be correct and $\tilde{w}_{ic|xh} = \tilde{\tilde{\pi}}_{c|xh}$ if it is determined to be incorrect.

The ML estimator $\tilde{\tilde{\pi}}_{c|x}$ is obtained by iterating between (14), (15) and (16) until convergence.

### 4.3 Pseudo-Maximum Likelihood (PML)

This section discusses an alternative to the CML, discussed in section 4.2, which is referred to as Pseudo-Maximum Likelihood (see Chambers and Skinner 2003). It is essentially a weighting approach, which may be easier to implement than CML, and relies on the factorisation given in section 4.2. It involves solving weighted versions of the score functions, $\text{Score}(\boldsymbol{\pi}_x; \mathbf{d}) = \mathbf{0}_{C-1}$ and $\text{Score}(\boldsymbol{\beta}; \mathbf{d}) = \mathbf{0}_{\mathrm{K}}$ for $\boldsymbol{\pi}_x$ and $\boldsymbol{\beta}$ respectively, where a record's weight equals the inverse of the probability that the record will remain unlinked. We denote the probability that record $i$ will not remain unlinked by $t_i = E(\gamma_i)$ so that the unit weights are given by $q_i = t_i^{-1}$, where here $i = 1, ..., n^*$. Consequently the PML estimator for $\pi_{c|x}$ is

$$\tilde{\pi}^{\text{PML}}_{c|x} = \breve{n}_{c|x}\left(\sum_c \breve{n}_{c|x}\right)^{-1}, \quad (17)$$

where $\breve{n}_{c|xy} = \Sigma_{i\in U_l} q_i \tilde{w}_{ic|x}$. The estimate of $\tilde{\pi}^{\text{PML}}_{c|x}$ is obtained by iterating between updating $\tilde{w}_{ic|x}$, given by (7), and (17) until convergence. The PML estimator for $\boldsymbol{\beta}$ is the same as the ML estimator but where the estimating equation (5) now has unit weights of $q_i$. One possible approach to estimating the accuracy of the PML estimates under perfect linkage is to use the Bootstrap method as described earlier, but where now the weight $q_i$ is introduced.

To illustrate when unlinked records are not ignorable, consider linking a data base with personal employment status to another data base with education level. Also assume that age and sex variables, which are correlated with employment and education, are available on one of the data bases. After conducting a clerical review, we may find that records for young males are 50% more likely to remain unlinked than records for females. This could be because males are less likely to provide their personal information, which is useful in linkage. Clearly, records for males on the linked file need to be given a weight double that for females in order for joint analysis of employment status and educational level to be unbiased.

## 5. Empirical study

A quality study conducted by the Australian Bureau of Statistics involved linking the 2006 Census of Population and Housing to its Dress Rehearsal. The Census Dress Rehearsal collected information from 78,349 persons and was conducted one year before the Census. The 2006 Census collected information from more than 19 million people.

Within a short window, during which the 2006 Census data were being processed, name and address were available for both the Census and the Census Dress Rehearsal. During this time, the two files of person level records were linked using two different standards of information:

- *Gold Standard* (GS) used name, address, mesh block and selected Census data items. Mesh block is a geographic area typically containing 50 dwellings. All names and addresses were destroyed at the end of the Census processing period.
- *Bronze Standard* (BS) used mesh block and selected Census data items (*i.e.*, did not use name and address). This is a method proposed to be used for future linking work by the ABS.

Full details of the quality study and the linkage methodology are given in Solon and Bishop (2009). The role of GS in the quality study is critical. It provides a benchmark against which the reliability of BS can be compared. The usefulness of the GS as a benchmark is due to the fact that name and address are powerful variables for the purpose of identifying common individuals on the Census and CDR and that it was subjected to thorough clerical review. As a result, GS is assumed to correspond to perfect linkage. Accordingly, differences between estimates based on GS and BS are interpreted as error. In other words, interest focuses on the reliability of BS *relative* to GS.

### 5.1 Linking methodology

#### 5.1.1 Blocking and linking variables and the 1 – 1 assignment algorithm

This subsection provides an overview of the CDR-to-Census linkage methodology for BS. The linking method consisted of a sequence of passes, where each pass is

defined by a set of blocking and linking variables and a 1 - 1 assignment algorithm. In the case of multiple passes, only records not linked in the first pass are eligible to be linked in the second pass, and only records not linked in the second pass are eligible to be linked in the third pass, and so on.

Table 1 gives the blocking variables, denoted by "B" for the BS. For example, during Pass 1, a Census record and a CDR record are only considered as a possible link if they have the same value for mesh block.

Linking variables are used to measure the degree of agreement between a record pair. A high level of agreement suggests that the likelihood of the record pair constituting a correct link is high. Table 1 gives the linking variables, denoted by "L", for BS. For example, during Pass 1 of BS, a range of variables such as day, month and year of birth, country of birth and highest level of qualifications are used as linking variables.

**Table 1**
**An example of blocking (B) and linking (L) variables used when linking 2006 Census data with the Census Dress Rehearsal. Different blocking variables were used on each of the two passes**

| Variable | Pass 1 | Pass 2 |
|---|---|---|
| Day of birth | L | B |
| Month of birth | L | B |
| Year of birth | L | B |
| Sex | L | B |
| Indigenous status | L | L |
| Country of birth | L | L |
| Language spoken | L | L |
| Year of arrival | L | L |
| Marital status | L | L |
| Religious affiliation | L | L |
| Field of study of highest qualification | L | L |
| Level of highest qualification | L | L |
| Highest level of schooling | L | L |
| Mesh block | B | L |

An output from each pass is a score for all record pairs. The score is a measure of the level of agreement between the pair of records. We defer the formal definition of score (for details see (3.6), Conn and Bishop 2006) but illustrate how it can be interpreted below. Consider BS in Pass 2 where record pairs have the same full date of birth and sex; a record pair would be assigned a score of 23.5 if there is agreement on mesh block (+17) and year of arrival (+8) and disagreement on religion (–1.5) (in this example agreement status for other linking variables would contribute to the score but for illustration purposes we ignore them). The contribution to the score for agreement on mesh block (+17) is greater than that for agreement on year of arrival (+8) because the former is less likely to occur by chance alone.

To formalise the aim of the linkage algorithm, denote the score for record $i$ on the CDR and record $j$ on the Census during pass $p$ of BS by $r_{pij}$. The set of all record pair scores $r_{pij}$ and the cut-off $f_p$ were used by the linking package *Febrl* (see Christen and Churches 2005) to determine the optimal set of links in pass $p$. The term $f_p$ is the minimum value for the score in order for a record pair to be assigned as a link during pass $p$. The *Febrl* algorithm seeks to maximise $\sum_i r_{pij}$, subject to $r_{pij} > f_p$. Clearly, the number of links depends upon $f_p$.

In what follows, we evaluate BS with two different sets of cut-offs, where a set of cut-offs is defined by the pass 1 and 2 cut-offs. The first is referred to as the Very Low (VL) cut-off and is considered to be optimal cut-off since, for a range of cut-offs, its naive estimates were "closest" to the corresponding GS estimates (see Bishop 2009). The second cut-off is referred to as Ultra-Low (UL) and effectively seeks to maximise the number of linked CDR records. Below we refer to the two BS linked files by their cut-offs, VL and UL.

### 5.1.2 Linking results

GS linked 70,274 of the 78,349 CDR records. Under the assumption that GS corresponds to perfect linkage, there were 8,075 individuals with CDR records but no Census records. In reality the GS is not perfect. For a discussion on this see Bishop 2009.

VL linked 57,790 CDR records. Of the 70,274 CDR records that were linked by GS, 13,784 remained unlinked by VL, 700 were linked incorrectly by VL and 55,790 were linked correctly by VL. Also, 1,300 CDR records were linked by VL but were not linked by GS- these are also incorrect links. So in total there were 2,000 (= 700 + 1,300) incorrect links.

UL linked 74,350 CDR records. Of the 70,274 CDR records that were linked by GS, 2,811 remained unlinked by UL, 9,793 were linked incorrectly by UL and 57,670 were linked correctly by UL. Also, 6,887 CDR records were linked by UL but were not linked by GS.

In summary, 97% of the VL links are correct and 20% (= 13,784/70,274) of the GS' CDR records remain unlinked. The corresponding figures for UL are 78% and 4% (= 2,811/70,274).

### 5.1.3 Modelling the probability of a link being correct

*All* UL and VL links were known to be correct or incorrect (*e.g.*, if a UL link is also made by GS then the UL link is correct. Otherwise the UL link is incorrect). As a result, $p_{xy^*}$ in section 3.1 was known from GS. However, to simulate reality, $p_{xy^*}$ was estimated from a clerical sample of size 1,000 that was selected from the linked files by simple random sampling.

### 5.1.4 Modelling the probability of a record remaining unlinked

Each CDR record linked by the GS was assigned a variable which indicated whether the record was unlinked by BS. Namely, if the record remained unlinked by BS then the indicator variable was assigned a '1' otherwise a '0'. A logistic model was fitted using GS, where the response variable was the above indicator variable and the explanatory variables were obtained from the CDR. The more than 20 explanatory variables that are in the model were selected by standard forward-backward model selection. The explanatory variables included educational level, language, born overseas, Indigenous status, and indicators of missing key variables such as *meshblock*. The resulting prediction resulted in $t_i$ and was used below to implement the Pseudo-ML method for both contingency tables and logistic regression.

## 5.2 Results of tabular analysis

Table 2 gives the results of cross-tabulating employment status of indigenous people as reported on the CDR and Census. Table 2a shows that the GS estimate of the proportion of indigenous people employed in the Census, given they were employed in CDR, is 78.3%. The corresponding naive estimate for VL, which assumes the data are perfectly linked, is 86.7%. Even after replacing each of the 700 incorrect VL links by their corresponding correct link and discarding the 1,300 linked records for which no correct link exists, the naive estimate is largely unchanged at 86.0% (referred to as *Gold Links* in Table 2a). This shows that the difference between the VL and GS estimates is not so much due to incorrect links but is mainly due to unlinked records. This explains in part why the ML estimate (86.4%) for VL (see section 3.1), which only corrects for incorrect links, did not lead to much improvement. Conditional ML (CML) (see section 4) was considered in an attempt to reduce the error due to unlinked records that may have led to a misrepresentation, with respect to age and sex characteristics, in the linked file. The CML employment estimate was 86.6%. Unfortunately, CML did not make much of an improvement, indicating that the underlying mechanism generating unlinked records did not depend upon age and sex. PML estimates (see section 4) also did not make much of an improvement, indicating that the logistic model described in section 5.1.4 did not explain the mechanism generating unlinked records. Interestingly, the ML estimate using UL was 81.8%- by far the closest estimate to the GS estimate of 78.3%. The UL's main source of error is due to incorrect links, the type of linkage error which the ML estimator addresses. This indicates that correcting for errors due to incorrect links was much more successful than correcting for errors due to unlinked records.

Standard errors of the GS, naive and ML estimates are shown in parentheses in Table 2a. For VL and UL, ML standard errors are respectively about 25% and 75% larger than the corresponding naive standard errors. Also, the ML standard errors for UL are slightly smaller than for VL indicating that the extra links made by UL were worthwhile. Clearly, naive inference with UL over-states the level of confidence in estimates. For VL, naive and ML standard errors and estimates are very close.

Irrespective of the cut-off, the ML estimates in Table 2 a, b and c are always closer to the GS estimates than the corresponding naive estimate. For example in Table 2b the ML estimates for VL is 36.9%, noticeably closer to the GS estimate of 37.9% than the naive estimate of 33.3%. Based on the estimates in Table 2 it could be argued that the choice of whether to use VL or UL is not so important, as along as the ML estimator is used.

**Table 2**
**Percentages of Indigenous persons in various employment categories in 2006 given their employment category in 2005. For each linked data set, Very Low and Ultra Low, the estimation methods can be compared with the Gold**

| | Estimates for different methods and linked data set | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **a: Indigenous persons employed in 2005** | | | | | | | | |
| ***Status in 2006*** | **Gold** | | | **Very Low Cut-off** | | | **Ultra Low Cut-off** | |
| | | **Naive** | **Gold links** | **ML** | **PML** | **CML** | **Naive** | **ML** |
| *Employed* | 78.3 | 86.7 | 86.0 | 86.4 | 86.6 | 86.1 | 71.9 | 81.8 |
| | (1.7) | (2.4) | | (3.0) | | | (1.7) | (2.9) |
| *Unemployed* | 3.7 | 4.2 | 4.3 | 4.1 | 4.1 | 4.2 | 6.3 | 3.3 |
| | (0.84) | (1.2) | | (2.5) | | | (0.82) | (2.1) |
| *Not in the labour force* | 17.8 | 9.0 | 9.6 | 9.3 | 9.1 | 9.6 | 21.6 | 14.7 |
| | (1.6) | (2.4) | | (3.1) | | | (1.6) | (2.8) |
| **b: Indigenous persons unemployed in 2005** | | | | | | | | |
| ***Status in 2006*** | **Gold** | | **Very Low** | | | | **Ultra Low** | |
| | | | **Naive** | **ML** | | | **Naive** | **ML** |
| *Employed* | 27.5 | | 27.7 | 27.2 | | | 35.2 | 23.8 |
| *Unemployed* | 34.4 | | 38.9 | 36.4 | | | 32.3 | 38.0 |
| *Not in the labour force* | 37.9 | | 33.3 | 36.3 | | | 32.3 | 38.0 |
| **c: Indigenous persons not-in-the-labour force in 2005** | | | | | | | | |
| *Employed* | 13.7 | | 10.8 | 10.7 | | | 24.3 | 10.5 |
| *Unemployed* | 5.8 | | 7.6 | 7.4 | | | 6.3 | 5.8 |
| *Not in the labour force* | 80.4 | | 81.5 | 81.8 | | | 69.2 | 83.5 |

Table 3 is the same as Table 2 except that it describes analyses of linked records from all persons 15 and over rather than only Indigenous persons. Again the ML always makes an improvement for the UL, though this is not the case for VL. Table 4 gives the student status in 2006 for persons who were students in 2005. Again the ML generally makes the estimates closer to the corresponding Gold estimate, especially for UL.

**Table 3**
**Percentages of all persons aged over 15 in various employment categories in 2006 given their employment category in 2005. For each linked data set, Very Low and Ultra Low, the estimation methods can be compared with the Gold**

| | | Estimates for different methods and linked data set | | | |
|---|---|---|---|---|---|
| **Status in 2006** | **Gold** | **Very Low** | | **Ultra Low** | |
| | | **Naive** | **ML** | **Naive** | **ML** |
| **a: Persons employed in 2005** | | | | | |
| *Employed* | 91.8 | 92.2 | 92.6 | 89.7 | 92.4 |
| *Unemployed* | 1.8 | 1.7 | 1.6 | 1.9 | 1.6 |
| *Not in the labour force* | 6.2 | 6.1 | 5.6 | 8.3 | 5.8 |
| **b: Persons unemployed in 2005** | | | | | |
| *Employed* | 44.5 | 44.3 | 44.0 | 49.4 | 43.8 |
| *Unemployed* | 26.8 | 26.6 | 27.5 | 22.8 | 27.6 |
| *Not in the labour force* | 28.6 | 28.7 | 28.4 | 27.6 | 28.5 |
| **c: Persons not-in-the-labour force in 2005** | | | | | |
| *Employed* | 12.1 | 12.3 | 11.1 | 16.8 | 11.0 |
| *Unemployed* | 3.1 | 3.1 | 3.0 | 3.0 | 3.0 |
| *Not in the labour force* | 84.7 | 84.5 | 85.7 | 80.1 | 85.9 |

**Table 4**
**Student outcomes in 2006 for high school students in 2005**

| **Student Status in 2006** | **Gold** | **Very Low** | | **Ultra Low** | |
|---|---|---|---|---|---|
| | | **Naive** | **ML** | **Naive** | **ML** |
| *High School Student* | 79.3 | 79.3 | 79.6 | 77.4 | 79.6 |
| *Completed High School* | 14.0 | 14.3 | 13.7 | 14.7 | 14.1 |
| *Did not Complete High School* | 6.6 | 6.3 | 6.6 | 7.8 | 6.2 |

### 5.3 Simulation

The following simulation study illustrates the problems with naive analysis and the benefit of using the method outlined in this paper. Files X and Y in the simulation, each containing 2,000 records, are independently generated 400 times, where each generated file is denoted by X($r$) and Y($r$), and $r = 1, ..., 400$. Specifically, on X($r$) $x_i$ is randomly generated from the Bernoulli distribution with parameter 0.5. On Y($r$), $y_i$ is randomly generated from the Bernoulli distribution with parameter $\upsilon_i$, where $\upsilon_i = 1 / [1 + \exp(\beta_0 + \beta_1 x_i)]$, $\boldsymbol{\beta} = (\beta_0, \beta_1)', \beta_0 = -0.5, \beta_1 = 1.5$. The $r^{\text{th}}$ set of imperfectly linked data, $\mathbf{d}^*(r)$, is generated by correctly linking each record on File Y($r$) to one record on File X($r$) with probability $p = 0.8, 0.90, 0.95$ and 1. For each $r^{\text{th}}$ set of linked data a clerical sample of 300 links is selected. Each link in the clerical sample is assigned as being correct or incorrect. We summarise the performance of the ML estimator from section 3.2.2 and the naive method, which assumes there is no linkage error, by their 95% coverage rates and their Mean Squared Error (MSE). The coverage rates are based on the standard errors calculated from the Bootstrap described in section 3.3 with $R = 40$ replicates. The MSE of $\tilde{\boldsymbol{\beta}}$ is calculated by

$$\text{MSE}(\tilde{\boldsymbol{\beta}}) = \frac{1}{400}\sum_{r=1}^{400}(\tilde{\boldsymbol{\beta}}_r - \boldsymbol{\beta})(\tilde{\boldsymbol{\beta}}_r - \boldsymbol{\beta})'$$

where $\tilde{\boldsymbol{\beta}}_r$ is the ML estimate of $\boldsymbol{\beta}$ from $\mathbf{d}^*(r)$.

Table 5 shows that the naive approach has poor coverage rates, due to its significant bias in the presence of linkage error, and consequently a relatively high MSE. The coverage rates for ML-Method 1 are very close to their nominal levels. The results show that, as the percentage of correct links reduces from 100% to 80%, the MSE of ML increases by a factor of about 3 for $\beta_0$ and $\beta_1$. (The coverage rates and MSE of ML Method 1and 2 were very similar so only the former are reported).

**Table 5**
**Mean squared error and coverage rates for linked simulated data, where correct linkage occurs with probability, p**

| | | **Mean Squared Error** | | | | **95% Coverage Rates** | | |
|---|---|---|---|---|---|---|---|---|
| | | 0.8 | 0.9 | 0.95 | 1 | 0.8 | 0.9 | 0.95 |
| Naive | $\beta_0$ | 0.024 | 0.010 | 0.0056 | 0.0043* | 0.35 | 0.80 | 0.93 |
| | $\beta_1$ | 0.11 | 0.038 | 0.016 | 0.011* | 0.05 | 0.62 | 0.88 |
| ML-Method 1 | $\beta_0$ | 0.013 | 0.0078 | 0.0055 | 0.0043* | 93.0 | 94.25 | 93.5 |
| | $\beta_1$ | 0.031 | 0.018 | 0.013 | 0.011* | 96.0 | 94.5 | 96.25 |

* when $p = 1$ the naive and ML estimators are the same by definition.

## 6. Discussion

Data linkage is an appropriate technique when data sets must be joined to enhance dimensions such as time and breadth or depth of detail. Data linkage is increasingly being used by statistical organisations around the world. It is well-known that errors can arise when linking files, for example when applying probabilistic linking methods. However, there has been little work reported in the literature about how to make valid inferences in the presence of such errors.

This paper provides methodological and practical advice to support analysts in this area.

In general, naively treating a linked file as if it were perfectly linked will lead to biased estimates. The analyst should only use the naive approach when both the number of unlinked records, defined as records that could be correctly linked but were not linked at all, and the number of incorrect links are negligible. This paper has presented a maximum likelihood approach to making valid inferences in the presence of both sources of error. The approach uses the well-known EM algorithm and is easy to apply in practice. The method can be applied when one of the files is not necessarily a subset of the other and when the linkage involves multiple passes. These situations often arise in practice, including many recent examples in the Australian Bureau of Statistics. The empirical study shows that the ML approach makes significant and meaningful improvements to the estimates from the linked data.

In the special case where File X is obtained by taking a random sample from File Y, the estimation procedure described is not 'full' maximum likelihood. This is because it does not use the fact that population totals for File Y are known. While inference using the method described here are still valid in this case, it could perhaps be made more efficient (see Scott and Wild 1997).

## Acknowledgements

The authors would like to thank Raymond Chambers and two reviewers from Survey Methodology for their contributions to this paper.

## References

Australian Bureau of Statistics (2008). Census Data Enhancement - Indigenous Mortality Quality Study, 2006-07. Information Paper catalogue no. 4723.0.

Bishop, G. (2009). Assessing the Likely Quality of the Statistical Longitudinal Census Dataset. Methodology Research Papers, catalogue no. 1351.0.55.026, Australian Bureau of Statistics, Canberra.

Chambers, R., Chipperfield, J.O., Davis, W. and Kovačević, M. (2009). Regression Inference Based on Estimating Equations and Probability-Linked Data. Submitted for publication.

Chambers, R.L., and Skinner, C.J. (2003). *Analysis of Survey Data*. New York: John Wiley & Sons, Inc.

Chambers, R. (2008). Regression analysis of probability-linked data. *Statisphere*, Volume 4, http://www.statisphere.govt.nz/official-statistics-research/series/vol-4.htm.

Christen, P., and Churches, T. (2005). Febrl – Freely extensible biomedical record linkage. Release 0.3.1, viewed 17 November 2008, http://cs.anu.edu.au/~Peter.Christen/Febrl/febrl-0.3/febrldoc-0.3/contents.html.

Conn, L., and Bishop, G. (2006). Exploring Methods for Creating a Longitudinal Census Dataset. Methodology Advisory Committee Papers, catalogue no. 1352.0.55.076, Australian Bureau of Statistics, Canberra.

Fair, M. (2004). Generalized record linkage system-Statistics Canada's record linkage software. *Austrian Journal of Statistics*, 33(1 and 2), 37-53.

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

Fuller, W.A. (1987). *Measurement Error Models*. New York: John Wiley & Sons, Inc.

Hausman, J.A., Abrevaya, J. and Scott-Morton, F.M. (1998). Misclassification of the dependent variable in a discrete-response setting. *Journal of Econometrics*, 87, 239-269.

Herzog, T.N., Scheuren, F.J. and Winkler, W.E. (2007). *Data Quality and Record Linkage Techniques*. New York: Springer.

Holman, C.D.J., Bass, A.J., Rouse, I.L. and Hobbs, M.S.T. (1999). Population-based linkage of health records in Western Australia: Development of a health services research linked database. *Australian and New Zealand Journal of Public Health*, 23(5), 453-459.

Lahiri, P., and Larsen, M.D. (2005). Regression analysis with linked data. *Journal of the American Statistical Association*, 100, 222-230.

National Center for Health Statistics (2009). Linkages between Survey Data from the National Center for Health Statistics and Program Data from the Social Security Administration. Methodology Report, http://www.cdc.gov/nchs/data/datalinkage/ssa_methods_report_2009.pdf.

Rubin, D.B., and Little, R.J.A. (2003). *Statistical analysis of missing data*, 2nd Edition. New York: John Wiley & Sons, Inc.

Scheuren, F., and Winkler, W.E. (1993). Regression analysis of data files that are computer matched. *Survey Methodology*, 19, 39-58.

Scott, A.J., and Wild, C.J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika*, 84, 57-71.

Solon, R., and Bishop, G. (2009). A Linkage Method for the Formation of the Statistical Longitudinal Census Dataset. Methodology Research Papers, catalogue no. 1351.0.55.025, Australian Bureau of Statistics, Canberra.

Winkler, W.E. (2001). Record Linkage Software and Methods for Merging Administrative Lists. Statistical Research Report Series, No. RR2001/03, Bureau of the Census.

Winkler, W.E. (2005). Approximate String Comparator Search Strategies for Very Large Administrative Lists. Statistical Research Report Series, no. RRS2005/02, Bureau of the Census.

Wright, J., Bishop, G. and Ayre, T. (2009). Assessing the Quality of Linking Migrant Settlement Records to Census Data. Methodology Research Papers, catalogue no. 1351.0.55.027, Australian Bureau of Statistics, Canberra.

Survey Methodology, June 2011
Vol. 37, No. 1, pp. 25-37
Statistics Canada, Catalogue No. 12-001-X



# Hierarchical Bayes small area estimation under a spatial model with application to health survey data

**Yong You and Qian M. Zhou** [1]

Abstract

In this paper we study small area estimation using area level models. We first consider the Fay-Herriot model (Fay and Herriot 1979) for the case of smoothed known sampling variances and the You-Chapman model (You and Chapman 2006) for the case of sampling variance modeling. Then we consider hierarchical Bayes (HB) spatial models that extend the Fay-Herriot and You-Chapman models by capturing both the geographically unstructured heterogeneity and spatial correlation effects among areas for local smoothing. The proposed models are implemented using the Gibbs sampling method for fully Bayesian inference. We apply the proposed models to the analysis of health survey data and make comparisons among the HB model-based estimates and direct design-based estimates. Our results have shown that the HB model-based estimates perform much better than the direct estimates. In addition, the proposed area level spatial models achieve smaller CVs than the Fay-Herriot and You-Chapman models, particularly for the areas with three or more neighbouring areas. Bayesian model comparison and model fit analysis are also presented.

Key Words: Area level model; Bayesian model comparison; Disease rate; Gibbs sampling; Hierarchical spatial model; Posterior predictive model checking; Sampling variance.

## 1. Introduction

Model-based small area estimation methods have been widely used in practice due to the increasing demand for precise estimates for local regions and various small areas. In general sample surveys are designed to provide reliable estimates for large regions or aggregates of small areas such as the whole nation and provinces. Direct survey estimates, based only on the area specific sample data, usually provide reliable estimates of the parameter of interest for those large areas. For small areas, particularly some small geographical areas or specific small domains, direct estimates are likely to yield large standard errors because of the small sample sizes in those small areas. Therefore in making inference for small areas, it is necessary to borrow strength from related areas to form indirect estimates that increase the effective sample size and thus increase the precision of estimates. It is now generally accepted that the indirect estimates should be based on explicit models that provide links to related areas through the use of supplementary data such as census counts or administrative records; see, for example, Rao (2003) and Jiang and Lahiri (2006) for more discussion on model-based small area methods. The model-based estimates are obtained to improve the direct design-based estimates in terms of precision and reliability, *i.e.*, smaller coefficients of variation (CVs). There are two broad classifications for small area models: area level models and unit level models. Area level models are based on area direct survey estimates and unit level models are based on individual observations in small areas. In this paper we focus on area level models that borrow strength across regions to improve the direct survey estimates.

Among the area level models, the Fay-Herriot model (Fay and Herriot 1979) is a basic and widely used area level model in practice to obtain reliable model-based estimates for small areas. The Fay-Herriot model basically has two components, namely, a sampling model for the direct estimates and a linking model for the parameters of interest. The sampling model involves the direct survey estimate and the corresponding sampling variance. The Fay-Herriot model assumes that the sampling variance is known in the model. Typically a smoothed estimator of the sampling variance is obtained and then treated as known in the model. Wang and Fuller (2003) and You and Chapman (2006) considered the situation where the sampling variances are unknown and modeled separately by direct estimators. In this paper we will consider both the smoothing and modeling methods for the sampling variances in the sampling model.

The linking model relates the parameter of interest to a regression model with area-specific random effects. In the Fay-Herriot model, the area random effects are usually assumed to be independent and identically distributed (*iid*) normal random variables to capture geographically unstructured variations among areas. However, in some small area applications, particularly in public health estimation problems, geographical variation of a disease is a subject of interest, and estimation of overall spatial pattern of risk and borrowing strength across regions to reduce variances of final estimates are both important. Thus, it may be more reasonable to construct spatial models on the area-specific

1. Yong You, Statistical Research and Innovation Division, Statistics Canada. E-mail: yongyou@statcan.gc.ca; Qian M. Zhou, Department of Biostatistics, Harvard University.

random effects to capture the spatial dependence among them. The spatial models are generally used in health related small area estimation, and various spatial models have been proposed for small area estimation (*e.g.*, Cressie 1990; Ghosh, Natarajan, Stroud and Carling 1998; Maiti 1998; Ghosh, Natarajan, Walter and Kim 1999; He and Sun 2000; Moura and Migon 2002; Singh, Shukla and Kundu 2005; Souza, Moura and Migon 2009). Best, Richardson and Thomson (2005) provided a comprehensive review on spatial models for disease mapping. Rao (2003) also discussed several spatial small area models.

The objective of this paper is to consider spatial correlation small area models and illustrate the usefulness of these models through an application to health survey data. The paper is organized as follows. In section 2, we first study area level models including the Fay-Herriot model and spatial correlation linking models. Then in section 3 we propose hierarchical Bayes (HB) small area models with spatial correlation and obtain HB inference for small area parameters through the Gibbs sampling method. In section 4, we apply the proposed models to the analysis of small area data from the Canadian Community Health Survey. We compare the performance of the model-based estimates with the direct design-based estimates, and moreover, we compare the proposed models with the Fay-Herriot model and the You-Chapman model (You and Chapman 2006) to investigate the effects of incorporating spatial structure on the area-specific random effects. Bayesian model comparison and model fit analysis are also provided. Finally in section 5, we offer some concluding remarks.

## 2. Small area models and inference

### 2.1 Fay-Herriot model

Let $\theta_i$ denote the parameter of interest for the $i^{\text{th}}$ area, where $i = 1, ..., m$, and $m$ is the total number of areas. The Fay-Herriot model assumes that the $\theta_i$'s are related to area specific auxiliary data $\mathbf{x}_i = (x_{i1}, ..., x_{ip})'$ through a linear regression model as follows:

$$\theta_i = \mathbf{x}_i'\beta + v_i,\ i = 1, ..., m \qquad (1)$$

where $\beta = (\beta_1, ..., \beta_p)'$ is the $p \times 1$ vector of regression coefficients, and the $v_i$'s are area-specific random effects assumed to be *iid* with $\mathrm{E}(v_i) = 0$ and $\mathrm{Var}(v_i) = \sigma_v^2$. The assumption of normality may also be included. This model is referred to as a linking model for $\theta_i$. The Fay-Herriot model also assumes that a direct survey estimator $y_i$, which is usually design-unbiased for the parameter of interest $\theta_i$, is available whenever the area sample size $n_i > 1$. It is customary to assume that

$$y_i = \theta_i + e_i,\ i = 1, ..., m \qquad (2)$$

where $e_i$'s are the sampling errors associated with the direct estimator $y_i$. We also assume that the $e_i$'s are independent normal random variables with mean $E(e_i \,|\, \theta_i) = 0$ and sampling variance $\mathrm{Var}(e_i \,|\, \theta_i) = \sigma_i^2$. The model (2) is referred to as a sampling model for the direct survey estimator $y_i$. Combining these two components (1) and (2) leads to a linear mixed effects model (the Fay-Herriot model) as

$$y_i = \mathbf{x}_i'\beta + v_i + e_i,\ i = 1, ..., m. \qquad (3)$$

In the basic Fay-Herriot model (3), the sampling variances $\sigma_i^2$ are usually assumed as known, which is a very strong assumption. Generally, we can use direct sampling variance estimates from the survey data, however, these direct estimates are unstable if sample sizes are small. Therefore, in practice, a smoothed estimator of $\sigma_i^2$ is used in the model and treated as known. A generalized variance function is usually applied in practice to obtain a smoothed estimator for the sampling variance, *e.g.*, Dick (1995). In recent years, a method of smoothing design effects has been developed and used in practice to obtain smoothed variance estimators (*e.g.*, Singh, Folsom and Vaish 2005; You 2008a; Liu, Lahiri and Kalton 2008). In particular, You (2008a) applied an equal design effects modeling approach to obtain smooth estimates of sampling variances. The design effect for the $i^{\text{th}}$ area may be approximately written as

$$\mathrm{deff}_i = \frac{s_i^2}{s_{ri}^2},\ \text{for } i = 1, ..., m,$$

where $s_i^2$ is the unbiased direct estimate of sampling variance based on the complex sampling design, and $s_{ri}^2$ is the estimate of sampling variance based on the assumption of simple random sampling design. For each area, based on the assumption of a common design effect, a smoothed factor deff can be obtained by $\mathrm{deff} = \sum_{i=1}^m \mathrm{deff}_i/m$. Then a smoothed sampling variance estimate $\tilde{\sigma}_i^2$ can be obtained as $\tilde{\sigma}_i^2 = s_{ri}^2 \cdot \mathrm{deff}$.

Instead of plugging in the smoothed estimates of sampling variances in the model, alternatively we can model the sampling variance directly. In the papers by Wang and Fuller (2003) and You and Chapman (2006), they assume the sampling variance $\sigma_i^2$ unknown and estimate $\sigma_i^2$ by an unbiased direct estimator $s_i^2$, which is independent of the direct survey estimator $y_i$. They also assume that $d_i s_i^2 \sim \sigma_i^2 \chi_{d_i}^2$, where $d_i = n_i - 1$, and $n_i$ is the sample size for the $i^{\text{th}}$ area. You and Chapman (2006) considered the full HB approach with the Gibbs sampling method which automatically takes into account the extra uncertainty associated with the estimation of $\sigma_i^2$. In this paper, we consider both the smoothing and modeling approaches for the sampling variances.

### 2.2 Spatial models

To incorporate spatially correlated random effects in the linking model, a simple and obvious way is to add a spatial random effect $u_i$ in the independent linking model (1) as follows:

$$\theta_i = \mathbf{x}_i'\beta + v_i + u_i, \qquad (4)$$

where $u_i$'s follow the well known intrinsic conditional autoregressive model given as

$$u_i \mid u_{-i} \sim \mathrm{N}\left( \frac{\sum_{j \neq i} w_{ij} u_j}{\sum_{j \neq i} w_{ij}}, \frac{\sigma_u^2}{\sum_{j \neq i} w_{ij}} \right), \qquad (5)$$

where $u_{-i}$ denotes the values of spatial random effects $u_j$'s in all other areas with $j \neq i$, weights $w_{ij}$ are fixed constants, and $\sigma_u^2$ is a unknown variance component. In practice, a common choice of $w_{ij}$ is to let $w_{ij} = 0$ unless areas $i$ and $j$ are neighboring areas (*i.e.*, share a common boundary), in which case $w_{ij} = 1$. The model (4) is proposed by Besag, York and Mollie (1991) to separate spatial effects from overall heterogeneity in the areas. In model (4), independent random effects $v_i$ capture geographically unstructured heterogeneity among areas, and spatial random effects $u_i$ capture spatial dependence between areas. In this way, the degree of overall spatial dependence can be expressed based on the proportion of the total variation in $v_i + u_i$ captured by each component.

In practice, it is often unclear how to choose between an unstructured model (*e.g.*, the basic linking model) given by (1) and a purely spatially structured model (*e.g.*, intrinsic autoregressive model) given by (5). For model (4), posterior inference about the spatial dependence is based on the proportion of the total variation in the sum of $v_i + u_i$ captured by each component. However, although the univariate conditional distributions of the spatial component (5) are well defined, the corresponding joint distribution is improper (with undefined mean and infinite variance). Moreover, the model (4) has a potential identifiability problem where only the sum of the random effects $v_i + u_i$ is well identified by the data; see, for example, Best *et al.* (2005), for a more detailed discussion.

Alternatively, we can consider another spatial parameterization studied by Leroux, Lei, and Breslow (1999) and MacNab (2003), which avoids the identifiability problem encountered with the model (4). Let $\theta_i = \mathbf{x}_i'\beta + b_i$, and $\mathbf{b} = (b_1, ..., b_m)'$. Following Leroux *et al.* (1999) and MacNab (2003), we place the following conditional autoregressive (CAR) model on the area specific spatial effects $\mathbf{b} = (b_1, ..., b_m)'$:

$$\mathbf{b} \sim \mathrm{MVN}(\mathbf{0}, \Sigma(\sigma_b^2, \lambda)) \qquad (6)$$

$$\Sigma(\sigma_b^2, \lambda) = \sigma_b^2 \mathbf{D}^{-1}, \ \mathbf{D} = \lambda \mathbf{R} + (1 - \lambda)\mathbf{I} \qquad (7)$$

where $\sigma_b^2$ is a spatial dispersion parameter and $\lambda$ is a spatial autocorrelation parameter, $0 \leq \lambda \leq 1$; $\mathbf{I}$ is an identity matrix of dimension $m$; $\mathbf{R}$, commonly known as the neighbourhood matrix, has $i^{\text{th}}$ diagonal element equal to the number of neighbors of the area $i$, and the off-diagonal elements in each row equal to -1 if the corresponding areas are neighbors and 0 otherwise. The CAR model (6) - (7) corresponds to the following conditional distribution of $b_i$:

$$b_i \mid b_{-i} \sim \mathrm{N}\left( \frac{\lambda}{1 - \lambda + \lambda w_{i+}} \sum_{j \neq i} w_{ij} v_j, \frac{\sigma_b^2}{1 - \lambda + \lambda w_{i+}} \right),$$

where $w_{i+} = \sum_{j \neq i} w_{ij}$. The CAR model (6) - (7) becomes the intrinsic autoregressive model (5) if $\lambda = 1$. On the other hand, if $\lambda = 0$, the CAR model (6) - (7) reduces to the independent linking model (1) which assumes independence on the area-specific random effects $v_i$. It is necessary to point out that the conditional mean and variances of $b_i \mid b_{-i}$ are weighted sums of the corresponding overall smoothing moments from the basic linking model (1) and local smoothing moments from the intrinsic autoregressive model:

$$\mathrm{E}(b_i \mid b_{-i}) = \frac{1 - \lambda}{1 - \lambda + \lambda w_{i+}} \times 0 + \frac{\lambda w_{i+}}{1 - \lambda + \lambda w_{i+}} \left( \sum_{j \neq} w_{ij} b_j / w_{i+} \right)$$

$$\mathrm{Var}(b_i \mid b_{-i}) = \frac{1 - \lambda}{1 - \lambda + \lambda w_{i+}} \times \sigma_b^2 + \frac{\lambda w_{i+}}{1 - \lambda + \lambda w_{i+}} (\sigma_b^2 / w_{i+}).$$

Thus model (6)-(7) is a balance between the independent linking model (1) and the intrinsic CAR model (5). The spatial correlation parameter $\lambda$ measures the extent of the spatial effects for local smoothing of the neighbouring areas. The modeling structure (6) captures both the unstructured heterogeneity among areas and the spatial correlation effects of the neighbouring area.

### 2.3 Hierarchical Bayes models and inference

In order to estimate $\theta_i$, the parameter of interest, we apply a hierarchical Bayes (HB) approach using the Gibbs sampling method. Compared to other approaches such as EBLUP and empirical Bayes (EB), HB approach is straightforward and the inference for $\theta_i$ are exact unlike the EB or EBLUP. Moreover, the HB approach can deal with complex small area models using the Monte Carlo Markov Chain

(MCMC) method, which overcomes the computational difficulties of multi-dimensional integrations of posterior quantities to a large extent.

Let $\mathbf{y} = (y_1, ..., y_m)'$, $\boldsymbol{\theta} = (\theta_1, ..., \theta_m)'$, and $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_m)'$. We first construct two HB models without and with spatial structure under the assumption that the sampling variance $\sigma_i^2$ are assumed known and replaced by the smoothed estimate $\tilde{\sigma}_i^2$.

*Model* 1: Fay-Herriot model, denoted as FHM (Fay and Herriot 1979; Rao 2003).

- $y_i \mid \theta_i \sim \mathrm{N}(\theta_i, \sigma_i^2 = \tilde{\sigma}_i^2)$, for $i = 1, ..., m$;
- $\theta_i \mid \beta, \sigma_v^2 \sim \mathrm{N}(\mathbf{x}_i'\beta, \sigma_v^2)$, for $i = 1, ..., m$;
- Priors for the parameters $(\beta, \sigma_v^2)$: $\pi(\beta) \propto 1$; $\pi(\sigma_v^2) \sim \mathrm{IG}(a_0, b_0)$, where $a_0, b_0$ are chosen to be very small known constants to reflect vague knowledge on $\sigma_v^2$. N stands for the normal distribution and IG for the inverse gamma distribution.

*Model* 2: Proposed area level CAR model, as an extension of the Fay-Herriot model, denoted as CAR-FHM.

- $\mathbf{y} \mid \boldsymbol{\theta} \sim \mathrm{MVN}(\boldsymbol{\theta}, \mathbf{E})$, where $\mathbf{E}$ is a diagonal matrix with the $i^{\text{th}}$ diagonal element $\sigma_i^2 = \tilde{\sigma}_i^2$;
- $\boldsymbol{\theta} \mid \beta, \sigma_v^2 \sim \mathrm{MVN}(\mathbf{X}\beta, \sigma_v^2\mathbf{D}^{-1})$, where $\mathbf{D} = \lambda\mathbf{R} + (1 - \lambda)\mathbf{I}$, with $\mathbf{I}$, an identity matrix of dimension $m$, and $\mathbf{R}$, the neighbourhood matrix;
- Priors for the parameters $(\beta, \lambda, \sigma_v^2)$: $\pi(\beta) \propto 1$; $\pi(\lambda) \sim \mathrm{Uniform}(0, 1)$, where $0 \le \lambda \le 1$; $\pi(\sigma_v^2) \sim \mathrm{IG}(a_0, b_0)$, where $a_0, b_0$ are chosen to be very small known constants. MVN stands for the multivariate normal distribution.

Note that the proposed model CAR-FHM reduces to FHM when the spatial autocorrelation parameter $\lambda = 0$.

We also consider two HB models with the sampling variance $\sigma_i^2$ unknown and modeled by the direct unbiased estimator $s_i^2$.

*Model* 3: You-Chapman Model, denoted as YCM (You and Chapman 2006).

- $y_i \mid \theta_i, \sigma_i^2 \sim \mathrm{N}(\theta_i, \sigma_i^2)$, for $i = 1, ..., m$;
- $d_i s_i^2 \mid \sigma_i^2 \overset{\text{ind}}{\sim} \sigma_i^2 \chi_{d_i}^2$ where $d_i = n_i - 1$, for $i = 1, ..., m$;
- $\theta_i \mid \beta, \sigma_v^2 \sim \mathrm{N}(\mathbf{x}_i'\beta, \sigma_v^2)$, for $i = 1, ..., m$;
- Priors for unknown parameters $(\beta, \sigma_v^2, \sigma_i^2, i = 1, ..., m)$: $\pi(\beta) \propto 1$; $\pi(\sigma_v^2) \sim \mathrm{IG}(a_0, b_0)$, $\pi(\sigma_i^2) \sim \mathrm{IG}(a_i, b_i)$ for $i = 1, ..., m$, where $a_i, b_i$ $(0 \le i \le m)$ are chosen to be very small known constants to reflect vague knowledge on $\sigma_i^2$ and $\sigma_v^2$.

*Model* 4: Proposed area level CAR model with unknown sampling variances, as an extension of You-Chapman model, denoted as CAR-YCM.

- $\mathbf{y} \mid \boldsymbol{\theta}, \sigma_1^2, ..., \sigma_m^2 \sim \mathrm{MVN}(\boldsymbol{\theta}, \mathbf{E})$, where matrix $\mathbf{E}$ has diagonal elements $\sigma_i^2$;
- $d_i s_i^2 \mid \sigma_i^2 \overset{\text{ind}}{\sim} \sigma_i^2 \chi_{d_i}^2$, where $d_i = n_i - 1$, for $i = 1, ..., m$;
- $\boldsymbol{\theta} \mid \beta, \sigma_v^2 \sim \mathrm{MVN}(\mathbf{X}\beta, \sigma_v^2 \mathbf{D}^{-1})$, where $\mathbf{D} = \lambda\mathbf{R} + (1 - \lambda)\mathbf{I}$;
- Priors for the parameters $(\beta, \lambda, \sigma_v^2, \sigma_i^2, i = 1, ..., m)$: $\pi(\beta) \propto 1$; $\pi(\lambda) \sim \mathrm{Uniform}(0, 1)$, where $0 \le \lambda \le 1$; $\pi(\sigma_v^2) \sim \mathrm{IG}(a_0, b_0)$; $\pi(\sigma_i^2) \sim \mathrm{IG}(a_i, b_i)$ for $i = 1, ..., m$, where $a_i, b_i$ $(0 \le i \le m)$ are chosen to be very small known constants.

Again, note that the proposed model CAR-YCM reduces to the You-Chapman model when $\lambda = 0$. For both models 3 and 4 there is an implicit assumption that the area-specific sample size $n_i \ge 2$. If flat priors are used for $\sigma_i^2$, we should have $n_i \ge 4$ to ensure proper posteriors (You and Chapman 2006).

We apply the Gibbs sampling method to estimate the posterior mean $\mathrm{E}(\theta_i \mid \mathbf{y})$ and the corresponding posterior variance $\mathrm{Var}(\theta_i \mid \mathbf{y})$. The required full conditional distributions of parameters under different models are given in Appendix A. For the Fay-Herriot model and the You-Chapman model, all the full conditional distributions have closed forms and drawing samples from these distributions is straightforward. For the proposed two area level spatial models CAR-FHM and CAR-YCM, the conditional distribution of the spatial correlation parameter $\lambda$ does not have a closed form. We use the Metropolis-Hastings algorithm within the Gibbs sampler (Chip and Greenberg 1995) to update $\lambda$. Under the model CAR-FHM, the full conditional distribution of $\lambda$ in the Gibbs sampler can be written as

$$[\lambda \mid \boldsymbol{\theta}, \beta, \sigma_v^2] \propto \mathrm{h}(\lambda)\,\mathrm{f}(\lambda)$$

where $\mathrm{f}(\lambda)$ is a density function of the uniform distribution, Uniform (0, 1), given as

$$\mathrm{f}(\lambda) \propto 1, \text{ where } 0 \le \lambda \le 1$$

and $\mathrm{h}(\lambda)$ is a function given by

$$\mathrm{h}(\lambda) \propto \left|\left[\lambda\mathbf{R} + (1 - \lambda)\mathbf{I}\right]^{-1}\right|^{-1/2} \times \exp\left\{-\frac{1}{2\sigma_v^2}(\boldsymbol{\theta} - \mathbf{X}\beta)'\left[\lambda\mathbf{R} + (1 - \lambda)\mathbf{I}\right](\boldsymbol{\theta} - \mathbf{X}\beta)\right\}.$$

We use $f(\lambda)$ as the "candidate" generating density function in the Metropolis-Hastings updating step. To update $\lambda$ from the current values of $(\boldsymbol{\theta}^{(k)}, \beta^{(k)}, \sigma_v^{2(k)})$, we proceed as follows:

1. Draw $\lambda^*$ from a uniform distribution;
2. Compute the acceptance probability $\alpha(\lambda^*, \lambda^{(k)}) = \min\{h(\lambda^*)/h(\lambda^{(k)}), 1\}$;
3. Generate $u$ from a uniform distribution, if $u < \alpha(\lambda^*, \lambda^{(k)})$, then the candidate value $\lambda^*$ is accepted, *i.e.*, $\lambda^{(k+1)} = \lambda^*$; otherwise $\lambda^*$ is rejected, and set $\lambda^{(k+1)} = \lambda^{(k)}$.

For the model CAR-YCM, a similar procedure can be applied when drawing samples from the conditional distribution of $\lambda$.

## 3. Data analysis

### 3.1 Data description and implementation

The Canadian Community Health Survey (CCHS) is a federal survey conducted by Statistics Canada. The primary objective of CCHS is to provide timely and reliable estimates of health determinants, health status and health system utilization across Canada. It is a cross-sectional survey which operates on a two-year collection cycle. The first year of the survey cycle "x.1" targets individuals aged 12 or older who are living in private dwellings, and it is a general population health survey with a large sample (130,000 persons) designed to provide reliable estimates at the health region, provincial and national levels. The second year of the survey cycle "x.2" has a smaller sample (30,000 persons) allocated based on provincial sample buy-ins and is designed to provide provincial and national level results on specific focused health topics. Although national and provincial estimates are very important, there is an increasing demand for health data at lower levels of geography voiced by a number of provinces including British Columbia (BC), Prince Edward Island (PEI), Quebec and others. Cycle "x.1" of the CCHS collected data corresponds to 136 health regions in the 10 provinces and three territories. It primarily used two sampling frames. The first one, used as the primary frame, was based on the area frame designed for the Canadian Labour Force Survey, and within the area frame, a multistage stratified cluster design was used to sample dwellings. The second frame consists of a list of telephone numbers. Random digit dialing methodology is used in some of the health regions for cost reasons. More details of the design are provided in Béland (2002). In this paper, we use a small data set from Cycle 1.1 as an example to demonstrate the analysis. We are interested in estimating the disease rate for local health regions within provinces. In particular, we apply the four models discussed in section 2 to estimate the asthma rate for 20 health regions in the province of BC using the data from Cycle 1.1. Figure 1 shows the map of the 20 health regions in the province of British Columbia. We use this map to define the neighbourhood correlation matrix used in the spatial models. Appendix B gives the list of health regions and related spatial structures.

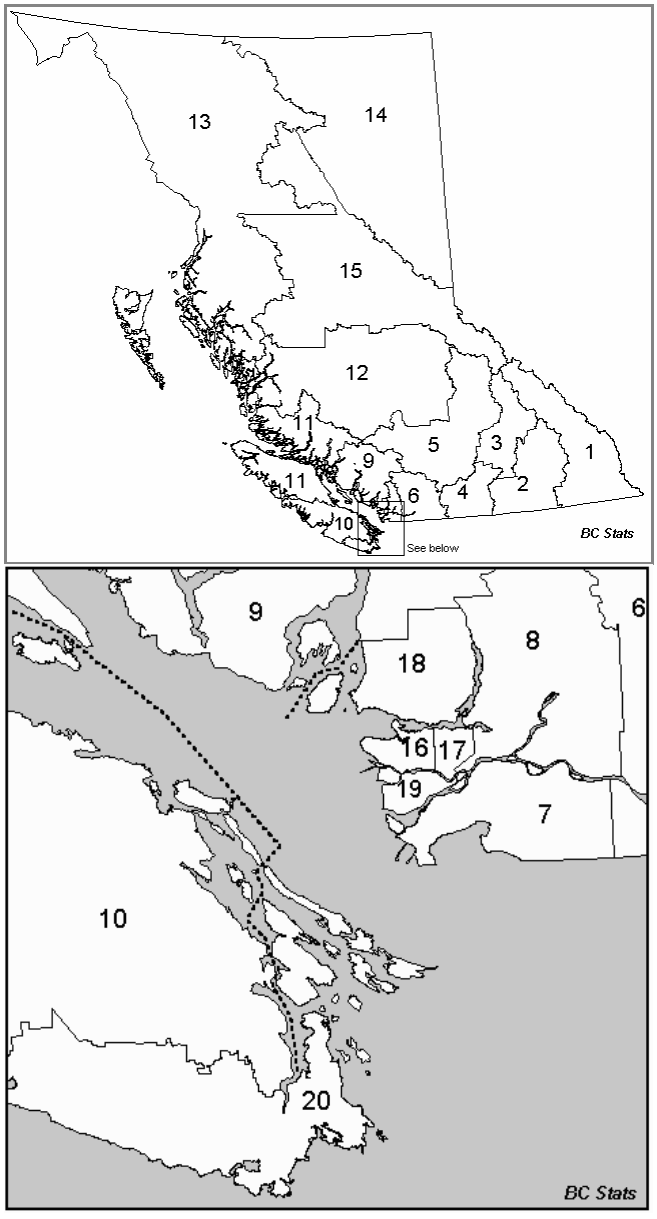


**Figure 1 Map of 20 health regions in the province of British Columbia**

Let $\theta_i$ denote the true asthma rate for the $i^{\text{th}}$ health region in BC, $i = 1, \ldots, 20$. From the survey data of Cycle 1.1, we obtained the direct survey estimate $y_i$ of $\theta_i$ as the ratio of number of people having asthma (direct survey

estimate) divided by the corresponding population size (known constant). We have also included six area level auxiliary variables used in the model, and these six variables are total population size, number of persons who have asthma as one of the symptoms of the chronic disease, number of persons who have asthma as the main symptom of the chronic disease, number of persons who have diabetes as one of the symptoms of the chronic disease, number of persons who have diabetes as the main symptom of the chronic disease, and number of visits to hospitals. Note that in the literature related to disease mapping (*e.g.*, Mollié 1996; Maiti 1998; MacNab 2003), a Poisson or Binomial distribution is usually assumed in the sampling model for the direct estimate $y_i$. However, in small area estimation, the direct estimate $y_i$ is obtained based on the complex sampling design used in the survey. Thus, it is a customary approach to assume a normal sampling model on the direct estimates $y_i$; see, for example, Datta, Lahiri, Maiti and Lu (1999), Rao (2003), Mohadjer, Rao, Liu, Krenzke and Van de Kerckhove (2007), and You (2008a) . Note that we have only considered one kind of disease rate data from one province in our study and used this example as illustration of the proposed model and evaluate the effects of spatial modeling in small area models.

To implement the Gibbs sampling, we use $L = 5$ parallel runs each with a "burn-in" length of $B = 2{,}000$ and Gibbs sampling size of $G = 5{,}000$. For the proposed models CAR-FHM and CAR-YCM, in order to reduce the autocorrelation which results from the accept-rejection algorithm in the run, we take every $5^{th}$ iteration after the "burn-in" period. Therefore, for models FHM and YCM, we have $n = 5{,}000$ samples for each run, and for models CAR-FHM and CAR-YCM, we have $n = 1{,}000$ samples for each run. Convergence of the Gibbs sampling is monitored for the small area parameters $\theta_i$ and other unknown parameters in the model using the potential scale reduction factor (Gelman and Rubin 1992; Gelman, Carlin, Stern and Rubin 2004, page 296-297). We have computed the reduction factors for all the monitored parameters in the model in the Gibbs sampling. These factor values are all very close to 1 (less than 1.05), which suggests that the desired convergence for these parameters is achieved by the Gibbs sampler.

We have used vague priors for the hyperparameters in the model as a common practice in HB small area estimation. In particular, the flat prior for regression parameter $\pi(\beta) \propto 1$ and proper inverse gamma priors for variance components are commonly used (*e.g.*, Arora and Lahiri 1997; Ghosh *et al.* 1998; Datta *et al.* 1999; You and Rao 2000; Rao 2003, page 237; Souza *et al.* 2009). Following MacNab (2003), we have used the uniform prior $\pi(\lambda) \sim \text{Uniform}(0, 1)$ for the autocorrelation parameter. The uniform priors are also commonly used for the autocorrelation parameters in spatial models (*e.g.*, Maiti 1998; He and Sun 2000; Rao 2003, page 266). We also tried several different values for the inverse gamma priors. The HB estimates are quite stable and not sensitive to the choice of vague proper priors. More detailed discussion on sensitivity analysis can be found, for example, in You and Chapman (2006) for similar models.

### 3.2 Comparison of results

At first, we present the HB estimates of the asthma rate under models FHM and CAR-FHM in which the sampling variances $\sigma_i^2$ are assumed to be known. We used the smoothed estimate $\tilde{\sigma}_i^2$ obtained by the smoothing technique in You (2008a) as described in Section 2. Figure 2 displays the direct estimates and the HB model-based estimates under FHM and CAR-FHM for the 20 health regions in BC. The health regions appear in the x-coordinate ranked by the order of sample size with the smallest (Peace Liard) on the left and the largest (South Fraser Valley) on the right. Model 1 (FHM) and Model 2 (CAR-FHM) give similar point estimates, and both the model-based estimates lead to moderate smooth estimates compared to the direct estimates. Moreover, the direct estimates and two HB estimates of the disease rate are very close for some health regions with large sample sizes, but for some areas with smaller sample sizes, they differ to some extent. Similar results are obtained under Model 3 (YCM) and Model 4 (CAR-YCM).

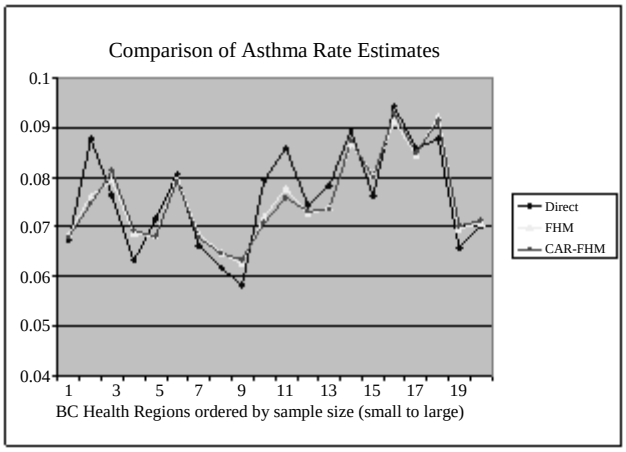


**Figure 2 Direct and HB model-based estimates under models FHM and CAR-FHM**

Figure 3 presents the CVs of the direct and two HB model-based estimates with the health regions ordered by the sample sizes from the smallest to the largest as in Figure 2. The CVs of HB estimates are obtained by dividing the squared root of the posterior variance by the posterior mean. As expected, the CVs of the direct estimates show a clear tendency of decrease as the sample size increases. However, the two model-based estimates give smoother CVs. Moreover, the two HB model-based estimates exhibit a great

improvement over the direct design-based estimates in terms of precision and reliability, that is, smaller CVs. Compared to the direct estimates, the average CV reduction of the HB estimates under FHM is about 22.7% ranging from 7.8% to 40.5%, and the average reduction of the CVs for the HB estimates under the proposed CAR-FHM is 27.8% ranging from 12.5% to 52.1%. Thus it is clear that the proposed spatial model CAR-FHM is superior to the Fay-Herriot model. We also obtained similar results for the models YCM and CAR-YCM when the sampling variance is modeled directly. The average CV reduction under YCM is 23.9%, whereas the average CV reduction is 29.0% under the proposed spatial model CAR-YCM. Details of the results including the point estimates and the corresponding CVs are presented in a table in Appendix C. In our example, the sample size at the health region level is relatively large. The model-based estimates have still shown great improvement over the direct survey estimates. Our results indicate that the presented small area models can be used to improve the direct survey estimates even when the sample size is relatively large. Note that Bayesian credible intervals for the small area parameters can be easily constructed using the MCMC output from the Gibbs sampler if required by practical users. This is an advantage of using the HB inference via MCMC sampling. However in this paper we only report the model-based point estimates and the corresponding CVs as our main purpose is to compare the model-based estimates with the direct estimates and to show the efficiency gain of the models. The gain in efficiency is clearly evident.

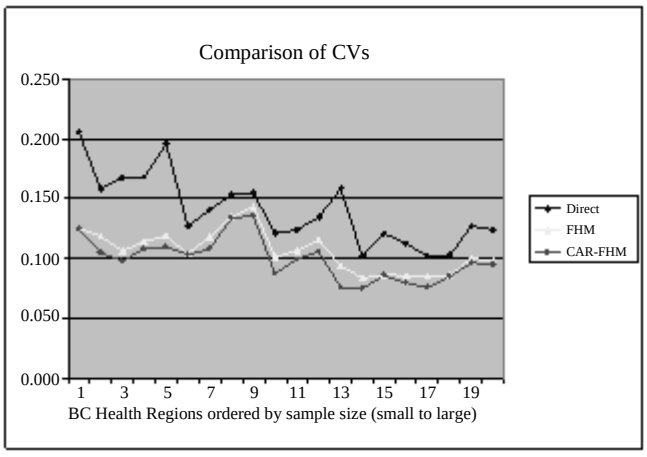


**Figure 3 Direct and HB CVs under models FHM and CAR-FHM**

In order to investigate the effects of incorporating the spatial structure in the model, we present the CVs of the direct and HB estimates by health regions sorted according to the number of neighbouring regions from the smallest (2 neighbours) to the largest (7 neighbours) in Figure 4. It shows that the HB estimates from the proposed model CAR-FHM has smaller CVs than the estimates from the Fay-Herriot model. In addition, the improvement of CAR-FHM over the Fay-Herriot model is much more obvious in the regions with more neighbours, and these two models give very close CVs in the regions with less adjacent areas. Very similar results are also obtained for CAR-YCM over YCM. Table 1 gives the average reduction of the CVs across the health regions with the same number of neighbours. The results in Table 1 present the CV reduction of the proposed spatial models for both cases of known and unknown sampling variances. For example, for known $\sigma_i^2$ (smoothed $\tilde{\sigma}_i^2$), for areas with only 2 neighbours, the average CV reduction of model CAR-FHM over the Fay-Herriot model is only around 0.9%, whereas for areas with 7 neighbours, the average CV reduction for CAR-FHM over FHM is as high as around 20%. For the case of unknown $\sigma_i^2$, similar results are obtained for CAR-YCM over YCM. The numerical results in Table 1 confirm the clear trend of increased CV reduction under the proposed spatial model over FHM or YCM as the number of neighbours increases. Thus, more neighbouring areas can provide more information in the spatial structure to improve the precision and reliability of the HB estimates.

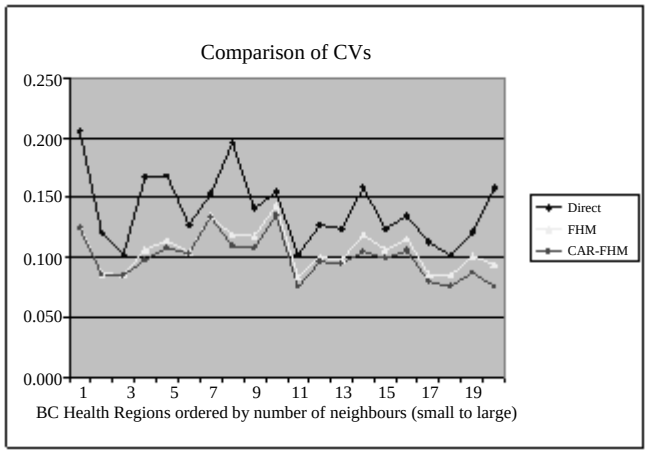


**Figure 4 Direct and HB CVs under models FHM and CAR-FHM with the health regions sorted by the number of neighbours**

**Table 1**
**Comparison of average CV reduction**

| Number of neighbours | Average CV reduction | |
|---|---|---|
| | CAR-FHM over FHM | CAR-YCM over YCM |
| 2 | 0.9% | 1.8% |
| 3 | 3.7% | 3.5% |
| 4 | 6.3% | 6.0% |
| 5 | 8.9% | 8.7% |
| 6 | 13.7% | 11.0% |
| 7 | 19.2% | 20.7% |

### 3.3 Bayesian model comparison

In this section, we compare the proposed models CAR-FHM with FHM and CAR-YCM with YCM, respectively. For hierarchical Bayes model comparison, the deviance information criterion (DIC) proposed by Spiegelhalter, Best, Carlin and van der Linde (2002) is commonly used in recent years to compare non-nested and mixed effects Bayesian models. The DIC is based on the deviance of the model $D(\theta)$, which is equal to minus twice the log-likelihood of the model, and the DIC is usually computed as DIC = $D(\hat{\theta}) + 2p_D$, where $D(\hat{\theta})$ is the deviance of the model evaluated at the posterior mean of the model parameters, which summarizes the goodness of fit of the model, and $p_D$ is the effective number of parameters, which captures the complexity of the model. $p_D$ is defined as $p_D = \bar{D}(\theta) - D(\hat{\theta})$, and $\bar{D}(\theta)$ is the posterior mean of the deviance of the model. Thus the DIC is defined as the summation of the goodness of fit of the model and the model complexity. Smaller values of DIC indicate a better model fit. Computation of DIC is relatively straightforward provided that the deviance $D(\theta)$ is available in closed form, and $p_D$ may be calculated after the Gibbs sampling run by taking the sample mean of the simulated values of $D(\theta)$ minus the plug-in estimate of the deviance $D(\hat{\theta})$. For the four models presented in section 2, we computed the corresponding DIC values, as shown in Table 2. It is clear that the proposed spatial models CAR-FHM and CAR-YCM both have smaller DIC values than the non-spatial models FHM and YCM respectively, which indicates that the spatial models are better than the non-spatial models in our study. Both spatial models CAR-FHM and CAR-YCM perform well in this example. This result of model comparison is consistent with the estimation results presented in section 3.2.

**Table 2**
**Comparison of DIC values for the four hierarchical models**

| Model | DIC value |
|---|---|
| FHM | 27.1 |
| CAR-FHM | 24.6 |
| YCM | 26.8 |
| CAR-YCM | 24.5 |

### 3.4 Test of model fit

In order to check the overall model fit of the proposed models CAR-FHM and CAR-YCM, we use the method of posterior predictive distribution. Let $y_{\text{rep}}$ denote the replicated observation under the model. The posterior predictive distribution of $y_{\text{rep}}$ given the observed data $y_{\text{obs}}$ is defined as $f(y_{\text{rep}} \mid y_{\text{obs}}) = \int f(y_{\text{rep}} \mid \theta)\, f(\theta \mid y_{\text{obs}})\, d\theta$. In this approach, a test statistic $T(y, \theta)$ that depends on the data $y$ and possibly the parameter $\theta$ can be defined and the observed value $T(y_{\text{obs}}, \theta \mid y_{\text{obs}})$ compared to the posterior predictive distribution of $T(y_{\text{rep}}, \theta \mid y_{\text{obs}})$ with any signifycant difference indicates a model failure. Lack of fit of the data with respect to the posterior predictive distribution can be measured by the *p*-value of the test quantity (Meng 1994; Gelman, Meng and Stern 1996). The posterior predictive *p*-value is defined as $p = P(T(y_{\text{rep}}, \theta) \geq T(y_{\text{obs}}, \theta) \mid y_{\text{obs}})$. If the given model adequately fits the observed data, then $T(y_{\text{obs}}, \theta \mid y_{\text{obs}})$ should be near the central part of the histogram of the $T(y_{\text{rep}}, \theta \mid y_{\text{obs}})$ values if $y_{\text{rep}}$ is generated repeatedly from the posterior predictive distribution. Consequently, the posterior predictive *p*-value is expected to be near 0.5 if the model adequately fits the data. Extreme *p*-values (near 0 or 1) suggest poor fit. The posterior predictive *p*-value model checking has been criticized for being conservative due to the double use of the observed data; see, for example, Bayarri and Berger (2000). They proposed alternative model checking *p*-value measures, named the partial posterior predictive *p*-value and the conditional predictive *p*-value. However, their methods are more difficult to implement and interpret (Rao 2003; Sinharay and Stern, 2003). As noted in Sinharay and Stern (2003), the posterior predictive *p*-value is especially useful if we think of the current model as a plausible ending point with modifications to be made only if substantial lack of fit is found.

To carry out the posterior predictive model checking, we need to specify a test quantity $T(y, \theta)$. You (2008b) studied several test quantities in posterior predictive model checking for small area models through a simulation study and proposed a test quantity given as

$$T(y, \theta) = \left| \max(y_i) - \text{mean}(\theta_i) \right| - \left| \min(y_i) - \text{mean}(\theta_i) \right|.$$

It is shown in You (2008b) that the proposed test quantity $T(y, \theta)$ is sensitive to the choice of distribution of random effects and different mean functions under the Fay-Herriot model. A similar test quantity is also suggested in Gelman *et al.* (2004) for posterior predictive model checking. In our study, under the proposed model CAR-FHM, the estimated *p*-value is 0.472, and under model CAR-YCM, the estimated *p*-value is 0.453. Thus there is no indication of lack of model fit and both proposed spatial models fit the data quite well.

To access model fit at the individual observation level, we also computed the individual predictive probability values $p_i^*$ as $p_i^* = P(y_{i(\text{rep})} < y_{i(\text{obs})} \mid y_{\text{obs}})$; see, for example, Gelfand (1996) and Daniels and Gatsonis (1999). These individual predictive probabilities provide information on the degree of consistent overestimation or underestimation of the observed data. For model CAR-FHM, the $p_i^*$ ranges from 0.325 to 0.768 with a mean of 0.517 and a median of 0.496; for model CAR-YCM, the $p_i^*$ ranges from 0.316 to

0.772 with a mean of 0.511 and a median of 0.497. Both models give very similar results and the mean and median values are all around 0.5. There is no indication of any consistent overestimation or underestimation of the proposed models. The overall *p*-values and individual predictive probabilities have shown that the proposed spatial small area models fit the data quite well.

### 3.5 Bias diagnostics

To evaluate any possible bias of the model-based estimates under the proposed models with respect to the direct survey estimates, following Brown, Chambers, Heady and Heasman (2001), we consider a simple method of regression analysis for the direct estimates and the HB model-based estimates. You (2008a) also used the regression analysis method for model bias diagnostics. If the model-based estimates are close to the true values of the small area disease rate, then the direct survey estimates, which are assumed to be unbiased for the true disease rates, should behave like random variables whose expected values correspond to the values of the model-based estimates. That means the model-based estimates should be unbiased predictors of the direct estimates. In terms of regression analysis, we basically fit the regression model $Y = \alpha + \beta X$ to the data and estimate the coefficients, and see how close the regression line is to $Y = X$. Let $Y$ be the direct survey estimates and $X$ be the model-based estimates. Under the proposed CAR-FHM, we obtain a regression line $Y = -0.0021(0.011) + 1.0365(0.1445)X$; under the proposed CAR-YCM, we obtain a regression line $Y = -0.0028(0.0108) + 1.0458(0.1427)X$. Thus both the regression lines show very little disparity from $Y = X$. We therefore conclude that the model-based estimates are consistent with the direct estimates with no extra possible bias induced by the proposed models. The results also provide an indication of no evidence of any bias due to possible model misspecification.

## 4. Conclusions

In this paper we have discussed two area level models, namely, the well-known Fay-Herriot model in which the sampling variance is assumed to be known, and the You-Chapman model in which the sampling variance is unknown and modeled separately by its direct estimator. In both the Fay-Herriot model and You-Chapman model, the area random effects are assumed to be *iid* normal random variables to capture unexplained area heterogeneity effects. After comparing various forms of Gaussian CAR models proposed in the literature (*e.g.*, Best *et al.* 2005) for disease mapping to incorporate spatially correlated effects, we extended the independent area effects model to a spatial correlation model and combined it with the traditional small area models. The proposed new small area spatial correlation models CAR-FHM and CAR-YCM include the small area sampling models and a spatial correlation linking model which captures both the unstructured heterogeneity among areas and the spatial correlation effects of the neighbouring areas. We don't need to specify the spatial autocorrelation parameter in the model, and this parameter will be estimated from the data.

In the data analysis we compared the proposed spatial models with the non-spatial effects models by applying the models to estimate the rates of asthma for 20 health regions in the province of British Columbia. Our results have shown that the model-based estimates achieve a great improvement over the direct estimates in terms of moderately smoothed point estimates and much smaller CVs. Particularly, the proposed models are superior to the Fay-Herriot model or You-Chapman model whether the sampling variances are assumed to be known or unknown. Moreover, note that the CV reduction of the proposed spatial models over the Fay-Herriot model or You-Chapman model is greater for the areas with more neighbours. Results of the Bayesian model comparison and model fit analysis are also in favor of the proposed small area spatial models.

In future work, the proposed small area spatial models can be extended to unmatched sampling and linking models (You and Rao 2002) with the sampling variance known or unknown. We plan to evaluate the estimation effects of different spatial models as well as the effects of spatial structures. For data analysis, we will produce model-based health status estimates based on the proposed models for health regions across Canada and evaluate the possibility of extending the model-based approach to lower level estimates such as age-sex domains within heath regions. We also plan to consider the data cloning method (Lele, Dennis and Lutscher 2007; Lele, Nadeem and Schmuland 2010) for the spatial models. An advantage of data cloning method is that the results are independent of the choice of priors. But the computational burden could be considerably extensive.

## Acknowledgements

We would like to thank one Associate Editor and one referee for their detailed comments and suggestions. Yong You's research work was supported by Statistics Canada Methodology Branch Research Block Fund. Qian M. Zhou's work was finished when she worked at Statistics Canada as a MITACS/NPCDS research internship student under the supervision of Yong You. Q.M. Zhou presented the proposed models and some results of the paper at the 2008 Statistics Society of Canada (SSC) annual meeting in Ottawa, and won the 2008 best student paper award of the SSC Survey Methods Section.

## Appendix A

## Full conditional distributions

A.1. Gibbs sampling full conditional distributions under Model 1: FHM.

- $[\theta_i \mid y_i, \beta, \sigma_v^2] \sim \mathrm{N}[\gamma_i y_i + (1-\gamma_i)\mathbf{x}_i'\beta, \tilde{\sigma}_i^2\gamma_i]$, where $\gamma_i = \sigma_v^2/(\sigma_v^2+\tilde{\sigma}_i^2)$, for $i = 1, ..., m$;
- $[\beta|\boldsymbol{\theta}, \sigma_v^2] \sim \mathrm{N}\left[\left(\sum_{i=1}^m \mathbf{x}_i\mathbf{x}'\right)^{-1}\left(\sum_{i=1}^m \mathbf{x}_i\theta_i\right), \sigma_v^2\left(\sum_{i=1}^m \mathbf{x}_i\mathbf{x}_i'\right)^{-1}\right]$;
- $[\sigma_v^2 \mid \boldsymbol{\theta}, \beta] \sim \mathrm{IG}\left[a_0 + \frac{1}{2}m, b_0 + \frac{1}{2}\sum_{i=1}^m (\theta_i - \mathbf{x}_i'\beta)^2\right]$.

A.2. Gibbs sampling full conditional distributions under Model 2: CAR-FHM.

- $[\boldsymbol{\theta} \mid \mathbf{y}, \beta, \lambda, \sigma_v^2] \sim \mathrm{MVN}(\boldsymbol{\Lambda}\mathbf{y} + (\mathbf{I}-\boldsymbol{\Lambda})\mathbf{X}\beta, \boldsymbol{\Lambda}\mathbf{E})$, where $\boldsymbol{\Lambda} = (\mathbf{E}^{-1}+\mathbf{D}/\sigma_v^2)^{-1}\mathbf{E}^{-1}$ with $\mathbf{E} = \mathrm{diag}\{\tilde{\sigma}_1^2, ..., \tilde{\sigma}_m^2\}$ and $\mathbf{D} = \lambda\mathbf{R} + (1-\lambda)\mathbf{I}$;
- $[\beta|\boldsymbol{\theta}, \lambda, \sigma_v^2] \sim \mathrm{MVN}[(\mathbf{X}'\mathbf{D}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}\boldsymbol{\theta}, \sigma_v^2(\mathbf{X}'\mathbf{D}\mathbf{X})^{-1}]$;
- $[\lambda|\boldsymbol{\theta}, \beta, \sigma_v^2] \propto \left|[\lambda\mathbf{R} + (1-\lambda)\mathbf{I}]^{-1}\right|^{-1/2} \times \exp\left\{-\frac{1}{2\sigma_v^2}(\boldsymbol{\theta} - X\beta)'[\lambda\mathbf{R} + (1-\lambda)\mathbf{I}](\boldsymbol{\theta} - X\beta)\right\}$;
- $[\sigma_v^2|\boldsymbol{\theta}, \beta, \lambda] \sim \mathrm{IG}\left[a_0 + \frac{m}{2}, b_0 + \frac{1}{2}(\boldsymbol{\theta} - X\beta)'\mathbf{D}(\boldsymbol{\theta} - X\beta)\right]$.

A.3. Gibbs sampling full conditional distributions under Model 3: YCM.

- $[\theta_i \mid y_i, \beta, \sigma_i^2, \sigma_v^2] \sim \mathrm{N}[\gamma_i y_i + (1-\gamma_i)\mathbf{x}_i'\beta, \sigma_i^2\gamma_i]$, where $\gamma_i = \sigma_v^2/(\sigma_v^2+\sigma_i^2)$, for $i = 1, ..., m$;
- $[\beta|\boldsymbol{\theta}, \sigma_v^2] \propto \mathrm{N}\left[\left(\sum_{i=1}^m \mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\sum_{i=1}^m \mathbf{x}_i\theta_i\right), \sigma_v^2\left(\sum_{i=1}^m \mathbf{x}_i\mathbf{x}_i'\right)^{-1}\right]$;
- $[\sigma_i^2 \mid y_i, \theta_i] \sim \mathrm{IG}\left(a_i + \frac{d_i+1}{2}, b_i + \frac{(y_i-\theta_i)^2 + d_i s_i^2}{2}\right)$, where $d_i = n_i - 1$, for $i = 1, ..., m$;
- $[\sigma_v^2 \mid \boldsymbol{\theta}, \beta] \sim \mathrm{IG}\left[a_0 + \frac{1}{2}m, b_0 + \frac{1}{2}\sum_{i=1}^m (\theta_i - \mathbf{x}_i'\beta)^2\right]$.

A.4. Gibbs sampling full conditional distributions under Model 4: CAR-YCM.

- $[\boldsymbol{\theta} \mid \mathbf{y}, \beta, \lambda, \sigma_v^2, \sigma_i^2] \sim \mathrm{MVN}(\boldsymbol{\Lambda}\mathbf{y} + (\mathbf{I}-\boldsymbol{\Lambda})\mathbf{X}\beta, \boldsymbol{\Lambda}\mathbf{E})$, where $\boldsymbol{\Lambda} = (\mathbf{E}^{-1}+\mathbf{D}/\sigma^2)^{-1}\mathbf{E}^{-1}$, and $\mathbf{E} = \mathrm{diag}\{\sigma_1^2, ..., \sigma_m^2\}$, $\mathbf{D} = \lambda\mathbf{R} + (1-\lambda)\mathbf{I}$;
- $[\beta \mid \boldsymbol{\theta}, \lambda, \sigma_v^2] \sim \mathrm{MVN}[(\mathbf{X}'\mathbf{D}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}\boldsymbol{\theta}, \sigma_v^2(\mathbf{X}'\mathbf{D}\mathbf{X})^{-1}]$;
- $[\lambda \mid \boldsymbol{\theta}, \beta, \sigma_v^2] \propto \left|[\lambda\mathbf{R} + (1-\lambda)\mathbf{I}]^{-1}\right|^{-1/2} \times \exp\left\{-\frac{1}{2\sigma_v^2}(\boldsymbol{\theta} - X\beta)'[\lambda\mathbf{R} + (1-\lambda)\mathbf{I}](\boldsymbol{\theta} - X\beta)\right\}$

  $[\sigma_i^2 \mid y_i, \theta_i] \sim \mathrm{IG}\left(a_i + \frac{d_i+1}{2}, b_i + \frac{(y_i-\theta_i)^2 + d_i s_i^2}{2}\right)$, where $d_i = n_i - 1$, for $i = 1, ..., m$;
- $[\sigma_v^2 \mid \boldsymbol{\theta}, \beta, \lambda] \sim \mathrm{IG}\left[a_0 + \frac{m}{2}, b_0 + \frac{1}{2}(\boldsymbol{\theta} - X\beta)'\mathbf{D}(\boldsymbol{\theta} - X\beta)\right]$.

# Appendix B

## List of 20 health regions in the province of British Columbia with the corresponding sample sizes and spatial structures

| ID number | Health region name | Sample size | Number of neighbours | Neighbours |
|---|---|---|---|---|
| 1 | East Kootenay | 645 | 3 | 2, 3, 15 |
| 2 | West Kootenay-Boundary | 705 | 3 | 1, 3, 4 |
| 3 | North Okanagan | 890 | 5 | 1, 2, 4, 5, 15 |
| 4 | South Okanagan Similameen | 1,063 | 4 | 2, 3, 5, 6 |
| 5 | Thompson | 982 | 7 | 3, 4, 6, 9, 11, 12, 15 |
| 6 | Fraser Valley | 1,125 | 5 | 4, 5, 7, 8, 9 |
| 7 | South Fraser Valley | 1,437 | 4 | 6, 8, 17, 19 |
| 8 | Simon Fraser | 1,165 | 5 | 6, 7, 9, 17, 18 |
| 9 | Coast Garibaldi | 623 | 5 | 5, 6, 8, 11, 18 |
| 10 | Central Vancouver Island | 1,077 | 2 | 11, 20 |
| 11 | Upper Island/Central Coast | 746 | 4 | 5, 9, 10, 12 |
| 12 | Cariboo | 673 | 4 | 5, 11, 13, 15 |
| 13 | North West | 650 | 3 | 12, 14, 15 |
| 14 | Peace Liard | 611 | 2 | 13, 15 |
| 15 | Northern Interior | 859 | 6 | 1, 3, 5, 12, 13, 14 |
| 16 | Vancouver | 1,285 | 4 | 17, 18, 19, 20 |
| 17 | Burnaby | 871 | 5 | 7, 8, 16, 18, 19 |
| 18 | North Shore | 842 | 4 | 8, 9, 16, 17 |
| 19 | Richmond | 828 | 3 | 7, 16, 17 |
| 20 | Capital | 1,225 | 2 | 10, 16 |

Note that Vancouver (#16) and Capital (#20) are not adjacent regions in the map since they are separated by the ocean. However, due to the intensive and close connection between these two regions, we define them as neighbours in our study for illustration purpose only.

# Appendix C

## Direct and model-based point estimates and CVs

| Comparison of point estimates | | | | | |
|---|---|---|---|---|---|
| **Area ID** | **Direct Est.** | **FHM** | **CAR-FHM** | **YCM** | **CAR-YCM** |
| 1 | 0.0765 | 0.0793 | 0.0812 | 0.0795 | 0.0812 |
| 2 | 0.0804 | 0.0795 | 0.0793 | 0.0797 | 0.0794 |
| 3 | 0.0745 | 0.0726 | 0.0731 | 0.0725 | 0.0729 |
| 4 | 0.0893 | 0.0868 | 0.0874 | 0.0867 | 0.0873 |
| 5 | 0.0782 | 0.0739 | 0.0736 | 0.0729 | 0.0731 |
| 6 | 0.0943 | 0.0914 | 0.0927 | 0.0918 | 0.0928 |
| 7 | 0.0702 | 0.0707 | 0.0712 | 0.0711 | 0.0717 |
| 8 | 0.0858 | 0.0845 | 0.0848 | 0.0844 | 0.0849 |
| 9 | 0.0877 | 0.0763 | 0.0745 | 0.0765 | 0.0747 |
| 10 | 0.0763 | 0.0805 | 0.0799 | 0.0805 | 0.0796 |
| 11 | 0.0661 | 0.0685 | 0.0678 | 0.0679 | 0.0676 |
| 12 | 0.0717 | 0.0681 | 0.0681 | 0.0678 | 0.0677 |
| 13 | 0.0631 | 0.0687 | 0.0692 | 0.0690 | 0.0693 |
| 14 | 0.0673 | 0.0685 | 0.0680 | 0.0685 | 0.0686 |
| 15 | 0.0793 | 0.0721 | 0.0707 | 0.0728 | 0.0713 |
| 16 | 0.0657 | 0.0696 | 0.0702 | 0.0697 | 0.0704 |
| 17 | 0.0859 | 0.0778 | 0.0759 | 0.0773 | 0.0759 |
| 18 | 0.0583 | 0.0626 | 0.0633 | 0.0618 | 0.0626 |
| 19 | 0.0619 | 0.0649 | 0.0647 | 0.0653 | 0.0647 |
| 20 | 0.0877 | 0.0923 | 0.0914 | 0.0917 | 0.0908 |
| **Comparison of CVs** | | | | | |
| **Area ID** | **Direct Est.** | **FHM** | **CAR-FHM** | **YCM** | **CAR-YCM** |
| 1 | 0.168 | 0.107 | 0.099 | 0.107 | 0.100 |
| 2 | 0.127 | 0.105 | 0.104 | 0.097 | 0.093 |
| 3 | 0.135 | 0.116 | 0.106 | 0.110 | 0.097 |
| 4 | 0.102 | 0.084 | 0.076 | 0.079 | 0.072 |
| 5 | 0.158 | 0.094 | 0.076 | 0.105 | 0.083 |
| 6 | 0.113 | 0.086 | 0.080 | 0.086 | 0.081 |
| 7 | 0.124 | 0.099 | 0.096 | 0.106 | 0.101 |
| 8 | 0.102 | 0.085 | 0.076 | 0.081 | 0.073 |
| 9 | 0.158 | 0.119 | 0.105 | 0.117 | 0.105 |
| 10 | 0.121 | 0.087 | 0.086 | 0.086 | 0.084 |
| 11 | 0.141 | 0.118 | 0.108 | 0.109 | 0.105 |
| 12 | 0.196 | 0.119 | 0.109 | 0.130 | 0.116 |
| 13 | 0.168 | 0.115 | 0.108 | 0.111 | 0.108 |
| 14 | 0.206 | 0.126 | 0.125 | 0.136 | 0.133 |
| 15 | 0.121 | 0.101 | 0.087 | 0.094 | 0.083 |
| 16 | 0.127 | 0.101 | 0.097 | 0.103 | 0.097 |
| 17 | 0.124 | 0.107 | 0.100 | 0.105 | 0.096 |
| 18 | 0.155 | 0.143 | 0.136 | 0.134 | 0.130 |
| 19 | 0.154 | 0.135 | 0.134 | 0.128 | 0.128 |
| 20 | 0.103 | 0.086 | 0.085 | 0.083 | 0.082 |

## References


Arora, V., and Lahiri, P. (1997). On the superiority of the Bayesian method over the BLUP in small area estimation problems. *Statistica Sinica*, 7, 1053-1063.

Bayarri, M.J., and Berger, J.O. (2000). P values for composite null models. *Journal of the American Statistical Association*, 95, 1127-1142.

Béland, Y. (2002). Canadian Community Health Survey Methodological Overview. Health Report, Statistics Canada, Catalogue no. 82-003, 13, 3, 9-14.

Besag, J., York, J. and Mollie, A. (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics*, 43, 1-59.

Best, N., Richardson, S. and Thomson, A. (2005). A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research*, 14, 35-39.

Brown, G., Chambers, R., Heady, P. and Heasman, D. (2001). Evaluation of small area estimation methods – An application to unemployment estimates from the UK LFS. Proceedings: Symposium 2001, *Achieving Data Quality in a Statistical Agency*: *A Methodological Perspective*, CD-ROM, 1-10.

Chip, S., and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49, 327-335.

Cressie, N. (1990). Small area prediction of undercount using the general linear model. Proceedings: Symposium 1990, *Measurement and Improving Data Quality*, Statistics Canada, 93-105.

Daniels, M.J., and Gatsonis, C. (1999). Hierarchical generalized linear models in the analysis of variations in health care utilization. *Journal of the American Statistical Association*, 94, 29-42.

Datta, G.S., Lahiri, P., Maiti, T. and Lu, K.L. (1999). Hierarchical Bayes estimation of unemployment rates for the states of the U.S. *Journal of the American Statistical Association*, 94, 1074-1082.

Dick, P. (1995). Modelling net undercoverage in the 1991 Canadian census. *Survey Methodology*, 21, 45-54.

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.

Gelfand, A.E. (1996). Model determination using sampling-based methods. In *Markov Monte Carlo in Practice* (Eds., W.R. Gilks, S. Richardson and D.J. Spiegelhalter), London: Chapman & Hall, 145-161.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2004). *Bayesian Data Analysis*, 2nd Edition. Chapman & Hall/CRC.

Gelman, A., Meng, X.L. and Stern, H.S. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica*, 6, 733-807.

Gelman, A., and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457-472.

Ghosh, M., Natarajan, K., Stroud, T.W.F. and Carlin, B.P. (1998). Generalized linear models for small area estimation. *Journal of the American Statistical Association*, 93, 273-282.

Ghosh, M., Natarajan, K., Walter, L.A. and Kim, D.H. (1999). Hierarchical Bayes GLMs for the analysis of spatial data: An application to disease mapping. *Journal of Statistical Planning and Inference*, 75, 305-318.

He, Z., and Sun, D. (2000). Hierarchical Bayes estimation of hunting success rates with spatial correlations. *Biometrics*, 56, 360-367.

Jiang, J., and Lahiri, P. (2006). Mixed model prediction and small area estimation (with discussion). *Test*, 15, 1-96.

Lele, S.R., Dennis, B. and Lutscher, F. (2007). Data cloning: Easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. *Ecology Letters*, 10, 551-563.

Lele, S.R., Nadeem, K. and Schmuland, B. (2010). Estimability and likelihood inference for generalized linear mixed models using data cloning. *Journal of the American Statistical Association*, 105, 1617-1625.

Leroux, B.G., Lei, X. and Breslow, N. (1999). Estimation of disease rates in small areas: A new mixed model for spatial dependence. In *Statistical Models in Epidemiology, the Environment and Clinical Trials*, (Eds., M.E. Halloran and D. Berry). New York: Springer Verlag, 135-178.

Liu, B., Lahiri, P. and Kalton, G. (2008). Hieararchical Bayes modeling of survey weighted small area proportions. Unpublished Manuscript.

Maiti, T. (1998). Hierarchical Bayes estimation of mortality rates for disease mapping. *Journal of Statistical Planning and Inference*, 69, 339-348.

Mohadjer, L., Rao, J.N.K., Liu, B., Krenzke, T. and van de Kerckhove, W. (2007). Hierarchical Bayes small area estimates of adult literacy using unmatched sampling and linking models. *Proceedings of the American Statistical Association, Section of Survey Method Research.*

Moura, F.A.S., and Migon, H.S. (2002). Bayesian spatial models for small area proportions. *Statistical Modelling*, 2, 3, 183-201.

MacNab, Y.C. (2003). Hierarchial Bayesian spatial modeling of small-area rates of non-rare disease. *Statistics in Medicine*, 22, 1761-1773.



Maiti, T. (1998). Hierarchical Bayes estimation of mortality rates for disease mapping. *Journal of Statistical Planning and Inference*, 69, 339-348.



Meng, X.L. (1994). Posterior predictive p value. *The Annals of Statistics*, 22, 1142-1160.

Mollié, A. (1996). Bayesian mapping of disease. In *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall, 359-379.

Rao, J.N.K. (2003). *Small Area Estimation*. New York: John Wiley & Sons, Inc.

Singh, A.C., Folsom, R.E., Jr. and Vaish, A.K. (2005). Small area modeling for survey data with smoothed error covariance structure via generalized design effects. Federal Committee on Statistical methods Conference proceedings, Washington, D.C., www.fcsm.gov.

Singh, B.B., Shukla, G.K. and Kundu, D. (2005). Spatio-temporal models in small area estimation. *Survey Methodology*, 31, 183-195.

Sinharay, S., and Stern, H.S. (2003). Posterior predictive model checking in hierarchical models. *Journal of Statistical Planning and Inference*, 111, 209-221.

Spiegelhalter, D.J., Best, N., Carlin, B.P. and van de Linde, A. (2002). Bayesain measures of model complexity and fit. *Journal of Royal Statistical Society*, B, 64, 583-639.

Souza, D.F., Moura, F.A.S. and Migon, H.S. (2009). Small area population prediction via hierarchical models. *Survey Methodology*, 35, 203-214.

Wang, J., and Fuller, W. A. (2003). The mean square error of small area predictors constructed with estimated area variances. *Journal of the American Statistical Association*, 98, 716-723.

You, Y. (2008a). An integrated modeling approach to unemployment rate estimation for sub-provincial areas of Canada. *Survey Methodology*, 34, 19-27.

You, Y. (2008b). Small area estimation using area level models with model checking and applications. *Proceedings of Survey Methods Section, Statistical Society of Canada.*

You, Y., and Chapman, B. (2006). Small area estimation using area level models and estimated sampling variances. *Survey Methodology*, 32, 97-103.

You, Y., and Rao, J.N.K. (2000). Hierarchical Bayes estimation of small area means using multi-level models. *Survey Methodology*, 26, 173-181.

You, Y., and Rao, J.N.K. (2002). Small area estimation using unmatched sampling and linking models. *The Canadian Journal of Statistics*, 20, 3-15.


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

Survey Methodology, June 2011
Vol. 37, No. 1, pp. 39-51
Statistics Canada, Catalogue No. 12-001-X



# Small area estimation under transformation to linearity

**Hukum Chandra and Ray Chambers** [1]

Abstract

Small area estimation based on linear mixed models can be inefficient when the underlying relationships are non-linear. In this paper we introduce SAE techniques for variables that can be modelled linearly following a non-linear transformation. In particular, we extend the model-based direct estimator of Chandra and Chambers (2005, 2009) to data that are consistent with a linear mixed model in the logarithmic scale, using model calibration to define appropriate weights for use in this estimator. Our results show that the resulting transformation-based estimator is both efficient and robust with respect to the distribution of the random effects in the model. An application to business survey data demonstrates the satisfactory performance of the method.

Key Words: Sample survey; Survey estimation; Business surveys; Model calibration; Skewed data; Model-based direct estimation; Empirical best linear unbiased prediction.

## 1. Introduction

Commonly used methods for small area estimation (SAE) assume that a linear mixed model can be used to characterize the regression relationship between the survey variable $Y$ and an auxiliary variable $X$ in the small areas of interest. In particular, empirical best linear unbiased prediction (EBLUP), see Rao (2003, chapters 6 - 8) is typically based on a linear mixed model assumption. However, when the data are skewed, as is often the case in business surveys, the relationship between $Y$ and $X$ may not be linear in the original (raw) scale, but can be linear in a transformed scale, *e.g.*, the logarithmic (log) scale. In such cases we would expect estimation based on a linear mixed model for $Y$ to be inefficient compared with one based on a similar model for a transformed version of $Y$. See Hidiroglou and Smith (2005). The use of transformations in inference has a long history, see for example Carroll and Ruppert (1988, chapter 4). Recently, Chen and Chen (1996) and Karlberg (2000a) have investigated the use of a 'transform to linearity' approach for regression estimation of survey variables that behave non-linearly. However, to the best of our knowledge there has been no application of this idea in SAE, even though economic theory (and casual observation) suggests that regression relationships in business survey data are typically multiplicative, and hence linear in the log scale.

In this paper we extend the model-based direct (MBD) estimation ideas described in Chandra and Chambers (2005, 2009) to the situation where the linear mixed model underpinning SAE holds on the log scale, using weights derived via model calibration (Wu and Sitter 2001). In doing so, we note that our approach easily generalises to other monotone (*i.e.*, invertible) transformations. In contrast, extension of the EBLUP approach to where the data follow a linear mixed model under transformation is complicated. We also relax the usual normality assumption for the area effects in order to examine robustness with respect to this assumption.

In the following section we summarise the MBD approach to SAE under a linear mixed model. In section 3 we describe an alternative to the linear mixed model for skewed data which reduces to the linear mixed model under log transformation, and in section 4 we use a model-based perspective to motivate model calibrated estimation of population quantities where the underlying variable is linear after suitable transformation. In section 5 we bring these two ideas together, introducing the concept of a fitted value model derived from a linear mixed model in the transformed scale. We then use this fitted value model to specify survey weights for use in an MBD estimator in SAE. In section 6 we present empirical results from a number of simulation studies that contrast the proposed transformation-based MBD estimator with both the EBLUP and the 'usual' MBD estimator defined by fitting a linear mixed model to the data as well as with an indirect empirical predictor based on the same transformed scale linear mixed model. Section 7 concludes the paper with a discussion of outstanding issues.

Note that the approach taken in this article is model-based. Consequently all moments are evaluated with respect to a model for the population data. Also, all sample data are assumed to have been obtained via a non-informative sampling method, *e.g.*, probability sampling with inclusion probabilities defined by known model covariates.

1. Hukum Chandra, Indian Agricultural Statistics Research Institute, Library Avenue, PUSA Campus, New Delhi-110012, India. E-mail: hchandra@iasri.res.in; Ray Chambers, Centre for Statistical and Survey Methodology, University of Wollongong, Wollongong, NSW, 2522, Australia. E-mail: ray@uow.edu.au.

## 2. Model-based direct estimation for small areas

To start, we fix our notation. Let $U$ denote a population of size $N$ and let $\mathbf{y}_U$ denote the $N$-vector of population values of a characteristic $Y$ of interest. Suppose that our primary aim is estimation of the total $t_{Uy} = \sum_U y_j$ of these population values (or their mean $m_{Uy} = N^{-1}\sum_U y_j$). Let $\mathbf{X}$ denote a $p$-vector of auxiliary variables that are related, in some sense, to $Y$ and let $\mathbf{x}_U$ denote the corresponding $N \times p$ matrix of population values these variables. We assume that the individual sample values of $\mathbf{X}$ are known. The non-sample values of $\mathbf{X}$ may not be individually known, but are assumed known at some aggregate level. At a minimum, we know the vector of population totals $\mathbf{t}_{Ux}$ of the columns of $\mathbf{X}$.

Suppose that it is reasonable to assume that the regression of $Y$ on $\mathbf{X}$ in the population is linear, *i.e.*,

$$E(\mathbf{y}_U | \mathbf{x}_U) = \mathbf{x}_U\beta \text{ and } \mathrm{Var}(\mathbf{y}_U | \mathbf{x}_U) = \mathbf{v}_U \qquad (1)$$

where $\mathbf{v}_U$ is known up to a multiplicative constant. Given a sample $s$ of size $n$ from this population, we can partition

$$\mathbf{x}_U = \begin{bmatrix} \mathbf{x}_s \\ \mathbf{x}_r \end{bmatrix}$$

and

$$\mathbf{v}_U = \begin{bmatrix} \mathbf{v}_{ss} & \mathbf{v}_{sr} \\ \mathbf{v}_{rs} & \mathbf{v}_{rr} \end{bmatrix}$$

into their sample and non-sample components. Here $r = U - s$ denotes the population units that are not in sample. The vector of weights that defines the Best Linear Unbiased Predictor (BLUP) of $t_{Uy}$ is then (Royall 1976; Valliant, Dorfman and Royall 2000, section 2.4)

$$\begin{aligned} \mathbf{w}_s^{\mathrm{BLUP}} &= (w_j^{\mathrm{BLUP}};\ j \in s) \\ &= \mathbf{1}_s + \mathbf{H}_s'(\mathbf{t}_{Ux} - \mathbf{t}_{sx}) + (\mathbf{I}_s - \mathbf{H}_s'\mathbf{x}_s')\mathbf{v}_{ss}^{-1}\mathbf{v}_{sr}\mathbf{1}_r \end{aligned} \qquad (2)$$

where $\mathbf{H}_s = (\mathbf{x}_s'\mathbf{v}_{ss}^{-1}\mathbf{x}_s)^{-1}\mathbf{x}_s'\mathbf{v}_{ss}^{-1}$, $\mathbf{I}_s$ is the identity matrix of order $n$, $\mathbf{t}_{sx}$ is the vector of sample totals of $\mathbf{X}$ and $\mathbf{1}_s$ ($\mathbf{1}_r$) denotes a vector of ones of size $n$ ($N - n$).

We now assume that the target population $U$ of size $N$ can be partitioned into $D$ non-overlapping small areas or domains, each of size $N_i$, $i = 1, ..., D$, such that $N = \sum_{i=1}^{D} N_i$. Given a sample $s$ of size $n$ units is drawn from this population, we shall assume that a sub-sample $s_i$ of size $n_i$ units is drawn from area $i$, with $n = \sum_{i=1}^{D} n_i$. Note that we assume that all small areas are sampled and that there is at least one sample unit in each small area of interest.

As noted in section 1, linear mixed models are often used in SAE. Such models can be written in the form

$$\mathbf{y}_U = \mathbf{x}_U\beta + \mathbf{g}_U\mathbf{u} + \mathbf{e}_U \qquad (3)$$

where $\mathbf{u}$ is a random vector of so-called area effects, $\mathbf{e}_U$ is a population $N$-vector of random individual effects and $\mathbf{g}_U$ is a known matrix. In general, area effects are vector-valued, so $\mathbf{u}' = (\mathbf{u}_1'\,\mathbf{u}_2'\cdots\mathbf{u}_D')$ and $\mathbf{g}_U = \mathrm{diag}\{\mathbf{g}_i;\ i = 1, \ldots, D\}$, where $\mathbf{g}_i$ is of dimension $N_i \times q$. The area specific effects $\{\mathbf{u}_i;\ i = 1, \ldots, D\}$ are assumed to be independent and identically distributed realisations of a random vector of dimension $q$ with zero mean and covariance matrix $\Sigma_u$. Similarly, the scalar individual effects making up $\mathbf{e}_U$ are assumed to be independent and identically distributed realisations of a random variable with zero mean and variance $\sigma_e^2$, with area and individual effects mutually independent. The parameters $\theta = (\Sigma_u, \sigma_e^2)$ are typically referred to as the variance components of (3).

Given the values of the variance components, it is straightforward to see that (3) is just a special case of the general linear model (1) that underpins the BLUP weights (2). In particular, under (3)

$$\begin{aligned} \mathbf{v}_{ss} &= \mathrm{diag}\{\mathbf{v}_{iss};\ i = 1, \ldots, D\} \\ &= \mathrm{diag}\{\mathbf{g}_{is}\Sigma_u\,\mathbf{g}_{is}' + \sigma_e^2\mathbf{I}_{is};\ i = 1, \ldots, D\} \end{aligned} \qquad (4)$$

and

$$\begin{aligned} \mathbf{v}_{sr} &= \mathrm{diag}\{\mathbf{v}_{isr};\ i = 1, \ldots, D\} \\ &= \mathrm{diag}\{\mathbf{g}_{is}\Sigma_u\,\mathbf{g}_{ir}';\ i = 1, \ldots, D\}. \end{aligned} \qquad (5)$$

Here $\mathbf{g}_{is}$ and $\mathbf{g}_{ir}$ denote the restriction of $\mathbf{g}_i$ to sampled and non-sampled units in area $i$ respectively. Given estimated values $\hat{\theta} = (\hat{\Sigma}_u, \hat{\sigma}_e^2)$ of the variance components we can substitute these in (4) and (5) to obtain estimates $\hat{\mathbf{v}}_{ss}$ and $\hat{\mathbf{v}}_{sr}$ of $\mathbf{v}_{ss}$ and $\mathbf{v}_{sr}$ respectively, and therefore compute 'empirical' BLUP weights, or EBLUP weights for the population total of $Y$ as

$$\begin{aligned} \mathbf{w}_s^{\mathrm{EBLUP}} &= (w_{ij}^{\mathrm{EBLUP}};\ j \in s_i;\ i = 1, \ldots, D) \\ &= \mathbf{1}_s + \hat{\mathbf{H}}_s'(\mathbf{t}_{Ux} - \mathbf{t}_{sx}) \\ &\quad + (\mathbf{I}_s - \hat{\mathbf{H}}_s'\mathbf{x}_s')\hat{\mathbf{v}}_{ss}^{-1}\hat{\mathbf{v}}_{sr}\mathbf{1}_r \end{aligned} \qquad (6)$$

where $\hat{\mathbf{H}}_s = (\mathbf{x}_s'\hat{\mathbf{v}}_{ss}^{-1}\mathbf{x}_s)^{-1}\mathbf{x}_s'\hat{\mathbf{v}}_{ss}^{-1}$. Note that we now use a double index of $ij$ to differentiate between population units in different areas.

The MBD estimator for the mean $m_{iy}$ of $Y$ in area $i$ (Chandra and Chambers 2005, 2009) based on the EBLUP weights for the total (6) is simply the corresponding weighted average of the sample values of $Y$ in area $i$,

$$\hat{m}_{iy}^{\text{HJ-LinMBD}} = \left\{\sum_{j\in s_i} w_{ij}^{\text{EBLUP}}\right\}^{-1} \sum_{j\in s_i} w_{ij}^{\text{EBLUP}} y_{ij}. \quad (7)$$

Note that (7) is *not* the EBLUP for $m_{iy}$ under (3). This is (see Rao 2003, section 6.2.3)

$$\begin{aligned}\hat{m}_{iy}^{\text{HT-LinEBLUP}} &= \hat{E}\{m_{iy} \mid \mathbf{y}_{is}, \mathbf{x}_{is}, \mathbf{x}_{ir}\} \\ &= N_i^{-1}\left[\sum_{j\in s_i} y_j + \mathbf{1}'_{ir}\left\{\mathbf{x}_{ir}\hat{\beta} + \hat{\mathbf{v}}_{irs}\hat{\mathbf{v}}_{iss}^{-1}(\mathbf{y}_{is} - \mathbf{x}_{is}\hat{\beta})\right\}\right] \\ &= N_i^{-1}\Big[n_i\bar{y}_{is} + (N_i - n_i) \\ &\quad \left\{\bar{\mathbf{x}}'_{ir}\hat{\beta} + \bar{\mathbf{g}}'_{ir}\hat{\Sigma}_u\mathbf{g}'_{is}(\mathbf{g}_{is}\hat{\Sigma}_u\mathbf{g}'_{is} + \hat{\sigma}_e^2\mathbf{I}_{is})^{-1}(\mathbf{y}_{is} - \mathbf{x}_{is}\hat{\beta})\right\}\Big]. \quad (8)\end{aligned}$$

Here $\hat{E}$ denotes the expectation operator under (3) with unknown parameters replaced by estimates, $\mathbf{x}_{is}$ and $\mathbf{x}_{ir}$ are the matrices of sample and non-sample values of $\mathbf{X}$ in area $i$, $\mathbf{y}_{is}$ is the vector of sample values of $Y$ in the same area, $\hat{\beta}$ is the 'empirical' BLUE of $\beta$, $\hat{\mathbf{v}}_{irs}$ is the transpose of the estimated value of $\mathbf{v}_{isr}$ with $\hat{\mathbf{v}}_{iss}$ the corresponding estimate of $\mathbf{v}_{iss}$, see (4) and (5), and $\mathbf{1}_{ir}$ is a vector of ones of length $N_i - n_i$. Note that the last expression on the right hand side of (8) follows directly by substitution of (4) and (5), with $\bar{\mathbf{x}}_{ir}$ and $\bar{\mathbf{g}}_{ir}$ denoting the column vectors of order $p$ and $q$ defined by averaging the columns of $\mathbf{x}_{ir}$ and $\mathbf{g}_{ir}$ respectively. Like the EBLUP (8), the estimator (7) is a weighted function of all the sample values. Note that under random intercept specification of (3), (8) reduces to the expression (7.2.39) in Rao (2003, section 7.2).

Mean squared error (MSE) estimation for (8) is usually carried out using the theory described in Prasad and Rao (1990). Although this MSE estimator is somewhat complicated, it works well under (3). However, when (3) fails it can be misleading. It is also inadequate as an estimator of the repeated sampling MSE of (8), as has been pointed out by Longford (2007). In contrast, MSE estimation for (7) is quite straightforward. This is because if one treats the weights defining this estimator as fixed, then it is a linear estimator of a domain mean, and so its prediction variance $V_i$ under (1) can be estimated using well-known methods (see Royall and Cumberland 1978). Since in general the EBLUP weights for the total (6) are not 'locally calibrated' (*i.e.*, they do not reproduce the area $i$ mean $\bar{\mathbf{x}}_i$ of $\mathbf{X}$), (7) has a bias $B_i$ under (1). A simple plug-in estimate of this bias is the difference between (7) and $\bar{\mathbf{x}}'_i\hat{\beta}$. The final MSE estimator used with (7) is therefore defined by summing the estimate of $V_i$ and the square of this estimate of $B_i$. This method of MSE estimation has been empirically demonstrated to have good model-based as well as repeated sampling properties. See Chandra and Chambers (2005, 2009), Chambers and Tzavidis (2006), Chandra, Salvati and Chambers (2007) and Tzavidis, Salvati, Pratesi and Chambers (2008).

## 3. Small area estimation under transformation

In this section we extend the MBD approach to SAE when the underlying regression relationships are non-linear. In doing so, we shall focus on the important case where the population values of $Y$ follow a non-linear model in their original (raw) scale, but their logarithms can be modelled linearly. The extension to other 'transform to linear' models is straightforward.

Without loss of generality, suppose that both $Y$ and $X$ are scalar and strictly positive, with skewed population marginal distributions and clear evidence of non-linearity in their relationship, *e.g.*, as in many business surveys applications. Furthermore, a linear mixed model is appropriate for characterising how the regression of $\log(Y)$ on $\log(X)$ varies between the small areas. That is, for $i = 1, ..., D$; $j = 1, ..., N_i$ we have

$$l_{ij} = \log(y_{ij}) = \beta_0 + \beta_1\log(x_{ij}) + \mathbf{g}'_{ij}\mathbf{u}_i + e_{ij} \quad (9)$$

where $y_{ij}$ and $x_{ij}$ are the values of $Y$ and $X$ respectively for population unit $j$ in small area $i$, $\mathbf{g}_{ij}$ denotes a 'contextual' covariate of dimension $q$, $\mathbf{u}_i$ denotes a random effect for area $i$ also of dimension $q$ and $e_{ij}$ is a scalar individual random effect. As usual with this type of model, we assume that all random effects are normally distributed and mutually uncorrelated, with zero expected values, $\text{Var}(\mathbf{u}_i) = \Sigma_u$ and $\text{Var}(e_{ij}) = \sigma_e^2$. Here $\Sigma_u$ is the $q \times q$ matrix of covariances for the random effects. Note that $\text{Var}(l_{ij} \mid x_{ij}) = v_{ijj} = \mathbf{g}'_{ij}\Sigma_u\mathbf{g}_{ij} + \sigma_e^2$ and $\text{Cov}(l_{ij}, l_{ik} \mid x_{ij}, x_{ik}, \mathbf{g}_{ij}, \mathbf{g}_{ik}) = v_{ijk} = \mathbf{g}'_{ij}\Sigma_u\mathbf{g}_{ik}$ under (9).

Given sample values of $y_{ij}$, $x_{ij}$ and $\mathbf{g}_{ij}$, standard methods of estimation (*e.g.*, ML or REML, see Harville 1977) can be used to estimate the parameters of (9). Let $\hat{\Sigma}_u$ and $\hat{\sigma}_e^2$ denote the resulting estimates of the variance components of this linear mixed model. The estimate of $\beta = (\beta_0\ \beta_1)'$ is then

$$\hat{\beta} = \left(\sum_i \mathbf{d}'_{is}\hat{\mathbf{v}}_{iss}^{-1}\mathbf{d}_{is}\right)^{-1}\left(\sum_i \mathbf{d}'_{is}\hat{\mathbf{v}}_{iss}^{-1}\mathbf{l}_{is}\right) \quad (10)$$

where $\hat{\mathbf{v}}_{iss}$, $\mathbf{d}_{is}$ and $\mathbf{l}_{is}$ are the sample components of $\hat{\mathbf{v}}_i = [\hat{v}_{ijk}] = \mathbf{g}_i\hat{\Sigma}_u\mathbf{g}'_i + \hat{\sigma}_e^2\mathbf{I}_i$, $\mathbf{d}_i = [d_{ijk}] = [\mathbf{1}_i\ \log(\mathbf{x}_i)]$ and $\mathbf{l}_i = (l_{ij};\ j = 1, \ldots, N_i)$ respectively. Here $\mathbf{g}_i$ is the $N_i \times q$ matrix defined by the covariates $\mathbf{g}_{ij}$ in area $i$, $\mathbf{I}_i$ is the identity matrix of order $N_i$, $\mathbf{1}_i$ denotes a vector of ones of dimension $N_i$ and $\log(\mathbf{x}_i)$ denotes the vector of $N_i$ values of $\log(X)$ in area $i$.

Note that when the variance components $\Sigma_u$ and $\sigma_e^2$ are known, (10) is the BLUE for $\beta$. Consequently, $E(\hat{\beta}) \approx \beta$

and $\text{Var}(\hat{\beta}) \approx (\sum_i \mathbf{d}'_{is} \hat{\mathbf{v}}_{iss}^{-1} \mathbf{d}_{is})^{-1}$. Put $\hat{\phi}_i = (\hat{\phi}_{ij}) = \mathbf{d}_i \hat{\beta}$. Then $E(\hat{\phi}_i) \approx \mathbf{d}_i \beta$ and $\text{Var}(\hat{\phi}_i) = \mathbf{A}_i = [a_{ijk}] \approx \mathbf{d}_i (\sum_g \mathbf{d}'_{gs} \hat{\mathbf{v}}_{gss}^{-1} \mathbf{d}_{gs})^{-1} \mathbf{d}'_i$, where $a_{ijk} = \mathbf{d}'_{ij} \text{Var}(\hat{\beta}) \mathbf{d}_{ik} \to 0$ as $n \to \infty$.

Our aim is to use the log scale linear mixed model (9) for estimation of the small area means $m_{iy}$. In particular, we use model calibration (Wu and Sitter 2001) based on this model to develop sample weights for use in the MBD estimator (7) of this quantity.

## 4. Model calibrated weighting

Model calibration was introduced by Wu and Sitter (2001) as a model-assisted method of calibrated weighting when the underlying regression relationship is non-linear. Here we provide a model-based perspective on the method, as a precursor to using it for constructing weights for use in an MBD estimator in a similar situation.

Suppose that the underlying population model is non-linear, with the relationship between $Y$ and $\mathbf{X}$ in the population of form

$$E(y_j \mid \mathbf{x}_j) = h(\mathbf{x}_j; \eta) \text{ and } \text{Var}(y_j \mid \mathbf{x}_j) = \sigma_j^2. \quad (11)$$

Here $j = 1, \ldots, N$, $\eta$ (typically vector-valued) and $\sigma_j^2$ are unknown model parameters and the mean function $h(\mathbf{x}_j; \eta)$ is a known function of $\mathbf{x}_j$ and $\eta$. We also assume that population units are mutually uncorrelated given their respective values of $\mathbf{X}$. Note that (11) is quite general, and includes linear, non-linear, and generalized linear models as special cases. In this situation, Wu and Sitter (2001) define the model-calibrated estimator of the population total $t_{Uy}$ as $\hat{t}_y^{mc} = \sum_{j \in s} w_j^{mc} y_j$, where the vector of weights $\mathbf{w}_s^{mc} = (w_j^{mc})$ is chosen to minimise an appropriately chosen measure of the distance from $\mathbf{w}_s^{mc}$ to the vector of Horvitz-Thompson weights $\mathbf{w}_s^{\pi} = (\pi_j^{-1})$, subject to the model calibration constraints

$$\sum_{j \in s} w_j^{mc} = N$$

and (12)

$$\sum_{j \in s} w_j^{mc} h(\mathbf{x}_j; \hat{\eta}_\pi) = \sum_{j \in U} h(\mathbf{x}_j; \hat{\eta}_\pi)$$

with $\hat{\eta}_\pi$ a design consistent estimator of $\eta$. Note that unlike standard calibration, the constraints (12) require that we know the individual population values of $\mathbf{X}$. The key idea behind this approach is that provided (11) fits reasonably, then $y_j$ is (at least approximately) a linear function of its fitted value $h(\mathbf{x}_j; \hat{\eta}_\pi)$ under this model and so we can carry out linear estimation using these fitted values as auxiliary information.

A model-based perspective on model calibration can be developed as follows. Let $\hat{\eta}$ denote a 'model-efficient' estimator of $\eta$ in (11), *e.g.*, its maximum likelihood (ML) estimator, with associated fitted values $h(\mathbf{x}_j; \hat{\eta})$. In general, these fitted values will not be unbiased. They will also be correlated. However, there will still be a systematic relationship between the actual values of $Y$ and their corresponding fitted values that we can approximate. Although there is nothing to stop us looking at more complex approximations, a linear model for the relationship between the population values $y_j$ and the fitted values $\hat{y}_j = h(\mathbf{x}_j; \hat{\eta})$ seems a reasonable starting point. We therefore replace the non-linear model (11) by the linear model

$$E(y_j \mid \hat{y}_j) = \alpha_0 + \alpha_1 \hat{y}_j$$

and (13)

$$\text{Cov}(y_j, y_k \mid \hat{y}_j, \hat{y}_k) = \omega_{jk}.$$

We refer to (13) as the 'fitted value' model corresponding to (11). Let $\mathbf{J}_U$ denote the population 'design matrix' under (13), *i.e.*, $\mathbf{J}_U = [\mathbf{1}_U \ \hat{\mathbf{y}}_U]$, where $\mathbf{1}_U$ denotes the unit vector of size $N$ and $\hat{\mathbf{y}}_U = (\hat{y}_j;\ j = 1, \ldots, N)$, and put $\Omega_U = [\omega_{jk};\ j = 1, \ldots, N;\ k = 1, \ldots, N]$. We can then partition $\mathbf{J}_U$ and $\Omega_U$ according to sample ($s$) and non-sample ($r$) units as

$$\mathbf{J}_U = \begin{bmatrix} \mathbf{J}_s \\ \mathbf{J}_r \end{bmatrix}$$

and

$$\Omega_U = \begin{bmatrix} \Omega_{ss} & \Omega_{sr} \\ \Omega_{rs} & \Omega_{rr} \end{bmatrix},$$

and hence write down the weights that define the BLUP of $t_{Uy}$ under (13). These are the model-based model-calibrated weights

$$\mathbf{w}^{mbmc} = (w_j^{mbmc};\ j \in s)$$
$$= \mathbf{1}_s + \mathbf{H}'_{cm}(\mathbf{J}'_U \mathbf{1}_U - \mathbf{J}'_s \mathbf{1}_s) + (\mathbf{I}_s - \mathbf{H}'_{cm} \mathbf{J}'_s) \Omega_{ss}^{-1} \Omega_{sr} \mathbf{1}_r \quad (14)$$

where $\mathbf{H}_{mc} = (\mathbf{J}'_s \Omega_{ss}^{-1} \mathbf{J}_s)^{-1} \mathbf{J}'_s \Omega_{ss}^{-1}$. Clearly, these weights are model-calibrated since $\sum_{j \in s} w_j^{mbmc} = N$ and $\sum_{j \in s} w_j^{mbmc} \hat{y}_j = \sum_{j \in U} \hat{y}_j$. However, unlike the linear model EBLUP weights (2), they are *not* calibrated on $\mathbf{X}$. In practice, the components of $\Omega_U$ will not be known and will need to be estimated. When these estimates are substituted in (14), we obtain the empirical version $\mathbf{w}^{embmc}$ of these model-calibrated weights.

## 5. Model calibrated weighting for small area estimation

We now use model calibration based on the log scale linear mixed model (9) to obtain sample weights for use in the MBD estimator (7). From the development in the previous section it can be seen that this requires us to first specify a fitted value model (13) for $Y$ based on (9), *i.e.*, we need to calculate appropriate fitted values $\hat{y}_{ij}$ as well as estimates $\hat{\omega}_{ijk}$ of $\omega_{ijk} = \text{Cov}(y_{ij}, y_{ik} | x_{ij}, x_{ik}, \mathbf{g}_{ij}, \mathbf{g}_{ik})$ under (9). The sample weights to use in the MBD estimator (7) are then given by (14).

A simple method of defining fitted values $\hat{y}_{ij}$ under (9) is one where parameter estimates derived under this model are used to obtain predicted values on the log scale which are then back-transformed. Unfortunately, as is well known, this approach is biased. We therefore develop the first and second order moments of an appropriate bias-corrected fitted value model based on (9). Let $\mathbf{x}_s$ and $\mathbf{g}_s$ denote the sample values of $x_{ij}$ and $\mathbf{g}_{ij}$ respectively. Under (9),

$$E(y_{ij} | x_{ij}, \mathbf{g}_{ij}) = E\{e^{l_{ij}} | x_{ij}, \mathbf{g}_{ij}\} = e^{\phi_{ij} + v_{ijj}/2}$$

$$\neq E(e^{\hat{\phi}_{ij} + \hat{v}_{ijj}/2} | \mathbf{x}_s, \mathbf{g}_s) = E(\hat{y}_{ij} | x_{ij}, \mathbf{g}_{ij})$$

so the usual bias correction that makes use of the fact that the conditional distribution of $y_{ij}$ is lognormal is inadequate. Let $\hat{\eta}_{ij} = (\hat{\beta}, \hat{v}_{ijj})'$ be an estimate of $\eta_{ij} = (\beta, v_{ijj})'$ such that $E(\hat{\eta}_{ij} - \eta_{ij}) \approx 0$ for large $n$. Put $z(\eta_{ij}) = e^{\phi_{ij} + v_{ijj}/2}$. Using a second order Taylor series approximation we can write

$$z(\hat{\eta}_{ij}) \approx z(\eta_{ij}) + (\hat{\eta}_{ij} - \eta_{ij})' z^{(1)}(\eta_{ij}) + \frac{1}{2}(\hat{\eta}_{ij} - \eta_{ij})' z^{(2)}(\eta_{ij})(\hat{\eta}_{ij} - \eta_{ij})$$

and so

$$E\{z(\hat{\eta}_{ij})\} \approx z(\eta_{ij}) + \frac{1}{2} tr[E\{z^{(2)}(\eta_{ij})(\hat{\eta}_{ij} - \eta_{ij})(\hat{\eta}_{ij} - \eta_{ij})'\}].$$

Here

$$z^{(1)}(\eta_{ij}) = \left( \mathbf{d}'_{ij} e^{\phi_{ij} + v_{ijj}/2} \quad \frac{1}{2} e^{\phi_{ij} + v_{ijj}/2} \right)'$$

and

$$z^{(2)}(\eta_{ij}) = \begin{pmatrix} \mathbf{d}_{ij}\mathbf{d}'_{ij} e^{\phi_{ij} + v_{ijj}/2} & \frac{1}{2}\mathbf{d}_{ij} e^{\phi_{ij} + v_{ijj}/2} \\ \frac{1}{2}\mathbf{d}'_{ij} e^{\phi_{ij} + v_{ijj}/2} & \frac{1}{4} e^{\phi_{ij} + v_{ijj}/2} \end{pmatrix}$$

are the vector and matrix respectively containing the first and second order derivatives of $z(\eta_{ij})$ with respect to $\eta_{ij}$. Since the asymptotic covariance between ML (or REML) estimators of the fixed and variance components of a linear mixed model is zero (McCulloch and Searle 2001, chapter 2, pages 40 - 45), the covariance between $\hat{\beta}$ and $\hat{v}_{ijj}$ will be negligible. It follows that

$$tr[E\{z^{(2)}(\eta_{ij})(\hat{\eta}_{ij} - \eta_{ij})(\hat{\eta}_{ij} - \eta_{ij})'\}]$$

$$= tr[z^{(2)}(\eta_{ij}) E\{(\hat{\eta}_{ij} - \eta_{ij})(\hat{\eta}_{ij} - \eta_{ij})'\}]$$

$$\approx e^{\phi_{ij} + \frac{v_{ijj}}{2}} \left[ \mathbf{d}'_{ij} \left( \sum_g \mathbf{d}'_{gs} \hat{\mathbf{v}}_{gss}^{-1} \mathbf{d}_{gs} \right)^{-1} \mathbf{d}_{ij} + \frac{1}{4} \text{Var}(\hat{v}_{ijj}) \right]$$

$$= E(y_{ij} | x_{ij}, \mathbf{g}_{ij}) \left[ \hat{a}_{ij} + \frac{1}{4} \text{Var}(\hat{v}_{ijj}) \right]$$

where $\hat{a}_{ijj} = \mathbf{d}'_{ij} \hat{V}(\hat{\beta}) \mathbf{d}_{ij}$ and $\hat{V}(\hat{\beta}) = (\sum_i \mathbf{d}'_{is} \hat{\mathbf{v}}_{iss}^{-1} \mathbf{d}_{is})^{-1}$ is the usual estimator of $\text{Var}(\hat{\beta})$. Our fitted values are therefore defined by the second order bias corrected estimator of $E(y_{ij} | x_{ij}, \mathbf{g}_{ij})$,

$$\hat{y}_{ij} = h(\mathbf{d}_{ij}; \hat{\eta}_{ij}) = \hat{k}_{ij}^{-1} e^{\hat{\phi}_{ij} + \hat{v}_{ijj}/2} \tag{15}$$

where

$$\hat{k}_{ij} = 1 + \frac{1}{2}\left\{ \hat{a}_{ijj} + \frac{1}{4}\hat{V}(\hat{v}_{ijj}) \right\}$$

and $\hat{V}(\hat{v}_{ijj})$ is the estimated asymptotic variance of $\hat{v}_{ijj}$. Under ML and REML estimation of the variance components of (9), this estimated asymptotic variance is obtained from the inverse of the relevant information matrix. Note that the bias adjustment of Karlberg (2000a) is a special case of (15).

In order to use (14) to define model-based model-calibrated sample weights, we also need estimates of the second order moments of the population values of $Y$ given these fitted values. The conditional moments $\omega_{ijk}$ are a first order approximation to these moments. In particular, given normal random effects

$$\omega_{ijk} = e^{(\phi_{ij} + \phi_{ik}) + (v_{ijj} + v_{ikk})/2} (e^{v_{ijk}} - 1) \tag{16}$$

Our estimate $\hat{\omega}_{ijk}$ of $\omega_{ijk}$ is obtained by substituting $\hat{\phi}_{ij}$ and $\hat{v}_{ijk}$ for $\phi_{ij}$ and $v_{ijk}$ in (16).

The empirical model-based model-calibrated weights (14) corresponding to the fitted value model defined by (15) and (16) are

$$\mathbf{w}^{embmc} = (w_{ij}^{embmc};\ j \in s_i;\ i = 1, \ldots, D)$$

$$= \mathbf{1}_s + \hat{\mathbf{H}}'_{mc}(\mathbf{J}'_U \mathbf{1}_U - \mathbf{J}'_s \mathbf{1}_s) + (\mathbf{I}_s - \hat{\mathbf{H}}'_{mc} \mathbf{J}'_s) \hat{\Omega}_{ss}^{-1} \hat{\Omega}_{sr} \mathbf{1}_r. \tag{17}$$

Here $\mathbf{J}_U = [\mathbf{1}_U \ \hat{\mathbf{y}}_U]$, so

$$\mathbf{J}_U'\mathbf{1}_U - \mathbf{J}_s'\mathbf{1}_s = \begin{pmatrix} N-n \\ \sum_i \sum_{j\in r_i} \hat{y}_{ij} \end{pmatrix},$$

and $\hat{\mathbf{H}}_{mc} = (\mathbf{J}_s'\hat{\Omega}_{ss}^{-1}\mathbf{J}_s)^{-1}\mathbf{J}_s'\hat{\Omega}_{ss}^{-1}$. Also $\hat{\Omega}_{ss} = \text{diag}\{\hat{\Omega}_{iss}; i = 1, \ldots, D\}$ and $\hat{\Omega}_{sr} = \text{diag}\{\hat{\Omega}_{isr}; i = 1, \ldots, D\}$, where $\hat{\Omega}_{iss}$ and $\hat{\Omega}_{isr}$ are defined by the sample/non-sample decomposition of $\hat{\Omega}_i$. For example, when (9) corresponds to a random intercepts specification, $\hat{v}_{ijk} = \hat{\sigma}_u^2 + \hat{\sigma}_e^2 I(j = k)$ and so the components of $\hat{\Omega}_i$ are

$$\hat{\omega}_{ijk} = e^{\hat{\phi}_{ij}+\hat{\phi}_{ik}+\hat{\sigma}_u^2+\hat{\sigma}_e^2}[e^{\hat{\sigma}_u^2}\{1 + I(j = k)(e^{\hat{\sigma}_e^2} - 1)\} - 1].$$

The development so far has assumed normality of log-scale random effects. However, there is no good reason (beyond convenience) to assume that with skewed data these random area effects should be normal. One alternative, given a scalar area effect in (9), is to assume that the random effects in this model are drawn from the *gamma* family of distributions. From the properties of this distribution and using binomial and exponential expansions (ignoring higher order terms) we can show that $E(y_{ij} | x_{ij}, \mathbf{g}_{ij}) \approx e^{\phi_{ij}+v_{ijj}/2} = z(\eta_{ij})$ as in the normal case. This indicates that an MBD estimator based on the model-based model-calibrated weights (17) should be robust with respect to the distribution of the random effects in (9).

Finally, we consider definition of the MBD estimator itself. As noted in section 2, this estimator is just the weighted average of the sample *Y*-values in an area. However, use of such a weighted average pre-supposes that the weights are reasonably close to being 'locally calibrated on *N*', *i.e.*, when summed over the sample units in small area *i* we obtain a value that is not too different from the actual small area population size $N_i$. This property usually holds if the weights are the EBLUP weights for the total (6) defined by a linear mixed model for *Y*. It does not necessarily hold for the model-based model-calibrated weights (17). Consequently, we consider two specifications for the MBD estimator given these weights. The first, which we refer to as a 'Hájek specification', is just the weighted average (7), with weights defined by (17). The second, which we refer to as a 'Horvitz-Thompson specification', replaces the denominator in (7) by the actual value of $N_i$. That is, the two types of MBD estimator under model-based model-calibrated weighting that we consider are

$$\hat{m}_{iy}^{\text{HJ-TrMBD}} = \left\{\sum_{j\in s_i} w_{ij}^{embmc}\right\}^{-1} \sum_{j\in s_i} w_{ij}^{embmc} y_{ij} \quad (18)$$

and

$$\hat{m}_{iy}^{\text{HT-TrMBD}} = N_i^{-1} \sum_{j\in s_i} w_{ij}^{embmc} y_{ij}. \quad (19)$$

Alternatively we can adopt a prediction-based approach to obtain an alternative indirect predictor for the small area mean under the log-transformed model (9). Our approach extends that of Karlberg (2000a). In this case, assuming model (9) holds, we predict each nonsample *Y* in small area *i* and then sum these predictions. Note that we need to correct for bias following back-transformation to the raw scale when calculating these predicted values for the nonsample *Y*. Under model (9), the resulting empirical predictor for the mean $m_{iy}$ of *Y* in area *i* (denoted TrEP) can be defined as

$$\hat{m}_{iy}^{\text{TrEP}} = N_i^{-1}\left\{\sum_{j\in s_i} y_{ij} + \sum_{j\in r_i} \hat{y}_{ij}\right\}, \quad (20)$$

where $\hat{y}_{ij}$ is given by (15).

Estimation of the MSE of (18) and (19) is carried out in the usual way for MBD estimators, *i.e.*, via the MSE estimation approach described in section 2. Estimation of the MSE of (20) is not straightforward since this predictor is a non-linear function of *Y* values. We do not pursue this issue in this paper.

## 6. An empirical evaluation

In this section we provide empirical results on the comparative performances of five different methods of SAE. These are the two 'transformation-based' MBD estimators (18) and (19), both based on the model-based model-calibrated weights (17) and denoted by HJ-TrMBD and HT-TrMBD respectively; the log-transformation based predictor (20) under model (9), denoted TrEP, the 'standard' MBD estimator (7) based on the linear mixed model (3) and the empirical EBLUP weights for the total (6), which we denote by HJ-LinMBD to emphasise that it is a Hájek-type weighted mean based on weights derived under a linear mixed model; and the EBLUP (8) derived under the same linear mixed model, which we denote HT-LinEBLUP. Note that the MSEs for all three MBD estimators were estimated using the method described in section 2, while the MSE of HT-LinEBLUP was estimated using the method described in Prasad and Rao (1990). Note that we have not considered estimation of the MSE of TrEP.

Our empirical results are based on two types of simulation studies. The first type used model-based simulation to generate artificial population and sample data. That is, at each simulation population data were first generated under the model and a single sample was then taken from this simulated population by stratified simple random sampling without replacement with small area as strata. These data were then used to compare the performances of the different estimators. In section 6.1 we present the results from these model-based simulations. We carried out two

sets of model-based simulations. In the first set of simulations (Set A), we investigated the performance of these estimators given population data generated using the log-scale linear mixed model (9). In second set of simulations (Set B), we examined the robustness of these estimators to misspecification of this model. The second type of simulation study was design-based. In section 6.2 we describe design-based simulations. Here we evaluated these estimators in the context of repeated sampling from a real population using realistic sampling methods. That is, real survey data were first used to simulate a population, and this fixed population was then repeatedly sampled according to a pre-specified design. In particular, the sample design used was stratified random sampling with strata corresponding to the small areas of interest and with stratum allocations set to the small area sample sizes in the original datasets.

Four measures of estimator performance were computed using the various estimates generated in these simulation studies. They were the relative bias (RB) and the relative root mean squared error (RRMSE) of these estimates, together with the coverage rate and average width of the nominal 95 per cent confidence intervals based on them. In Tables 2 to 4 these measures are presented as averages over the small areas of interest.

### 6.1 The model-based simulation study

Model-based simulations are a common way of illustrating the sensitivity of an estimation procedure to variation in assumptions about the structure of the population of interest. Here we fixed the population size at $N = 15,000$ and randomly generated the small area population sizes $N_i$, $i = 1, ..., D = 30$ so that $\sum_i N_i = N$. We used an overall sample size of $n = 600$ with small area sample sizes set so that they were proportional to the corresponding small area population sizes. These area-specific population and sample sizes were kept fixed in all our simulations. The population and sample sizes are given in Table 1a.

**Table 1a**
**Area specific population ($N_i$) and sample ($n_i$) sizes for model-based simulation**

| **Area** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $N_i$ | 525 | 538 | 510 | 468 | 526 | 484 | 516 | 458 | 529 | 518 |
| $n_i$ | 21 | 22 | 20 | 19 | 21 | 19 | 21 | 19 | 21 | 21 |
| **Area** | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| $N_i$ | 502 | 524 | 509 | 484 | 487 | 459 | 542 | 498 | 512 | 500 |
| $n_i$ | 20 | 21 | 20 | 19 | 19 | 18 | 22 | 20 | 20 | 20 |
| **Area** | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 |
| $N_i$ | 497 | 492 | 443 | 506 | 513 | 536 | 506 | 495 | 463 | 460 |
| $n_i$ | 20 | 20 | 18 | 20 | 21 | 21 | 20 | 20 | 19 | 18 |

In Set A of our model-based simulations the population values $y_{ij}$ were generated using the multiplicative model $y_{ij} = 5.0 x_{ij}^{\beta} u_i e_{ij} (j = 1,..., N_i;\ i = 1, ..., 30)$, with random samples then taken from each small area. Here the values of $x_{ij}$ were independently drawn from the log-normal distribution $\log(x_{ij}) \sim N(6, \sigma_x^2)$, with the individual effects and area effects independently drawn as $\log(e_{ij}) \sim N(0, \sigma_e^2)$ and $\log(u_i) \sim N(0, \sigma_u^2)$ respectively. The population values of $x$ were re-generated in each simulation. In particular, in each simulation we first generated the values of $x$'s for a population of size $N$ and then randomly assigned these values to different areas of sizes $N_i$. The values of $\sigma_e$ and $\sigma_u$ were chosen so that the intra-area correlation in the population varied between 0.20 and 0.25. Table 1b shows the six different sets of parameter values that were used in Set A. These ensured that the simulated populations contained a wide range of variation. For each generated population and for each area $i$ we selected a simple random sample (without replacement) of size $n_i$, leading to an overall sample size of $n = 600$. The sample values of $y$ and the population values of $x$ obtained in each simulation were then used to estimate the small area means. That is, using the sample data in each case, parameter values were estimated using the *lme* function in R (Bates and Pinheiro 1998), and estimates for the small area means then calculated, along with appropriate nominal 95% confidence intervals. The process of generating population and sample data, estimation of parameters and calculation of small area estimates was independently replicated 1,000 times. The results from this part of the simulation study are shown in Table 2.

**Table 1b**
**Population specifications for model-based simulation Set A**

| **Parameter Set** | $\beta$ | $\sigma_u$ | $\sigma_e$ | $\sigma_x$ |
|---|---|---|---|---|
| 1 | 0.5 | 0.30 | 0.50 | 3.00 |
| 2 | 0.8 | 0.35 | 0.60 | 2.50 |
| 3 | 1.0 | 0.40 | 0.70 | 2.25 |
| 4 | 1.3 | 0.45 | 0.80 | 1.75 |
| 5 | 1.5 | 0.50 | 0.90 | 1.50 |
| 6 | 2.0 | 0.60 | 1.00 | 1.20 |

In Set B of the model-based simulations, population data were generated using the model $y_{ij} = 5.0 x_{ij} [\exp(\log^2(x_{ij}))]^{\gamma} u_i e_{ij}$. Here the individual effects $e_{ij}$ and the area effects $u_i$ were independently drawn as $\log(e_{ij}) \sim N(0, 1)$ and $\log(u_i) \sim N(0, 0.25)$ respectively, while the covariate values $x_{ij}$ were drawn as $\log(x_{ij}) \sim N(3, 0.04)$. Five different values for the parameter $\gamma$ (-1.0, -0.5, 0.0, 0.5, 1.0) were investigated, thus generating population data with different degrees of curvature. All other aspects of these simulations, including the estimators considered, were the same as in Set A. Table 3 presents results from this component of the simulation study.

**Table 2**
**Average relative bias (ARB), average relative RMSE (ARRMSE), average coverage rate (ACR) and average interval width (AW) for model-based simulation Set A**

| Criterion | Estimator | Parameter Set | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 |
| ARB,% | HJ-TrMBD | -82.68 | -95.02 | -98.08 | -98.50 | -98.29 | -99.00 |
| | HT-TrMBD | 0.09 | 0.10 | -0.14 | -0.25 | -0.03 | 0.04 |
| | TrEP | 0.08 | 0.09 | -0.18 | -0.48 | -0.05 | 0.01 |
| | HJ-LinMBD | 12.01 | 4.09 | -1.35 | -5.54 | -6.60 | -9.88 |
| | HT-LinEBLUP | 13.39 | 5.18 | -0.67 | -5.24 | -6.41 | -9.67 |
| ARRMSE | HJ-TrMBD | 4.80 | 1.39 | 1.25 | 1.44 | 1.42 | 1.62 |
| | HT-TrMBD | 0.15 | 0.26 | 0.45 | 0.64 | 0.66 | 0.91 |
| | TrEP | 0.30 | 0.41 | 0.58 | 0.80 | 0.81 | 1.09 |
| | HJ-LinMBD | 1.11 | 1.41 | 1.85 | 1.99 | 2.06 | 2.69 |
| | HT-LinEBLUP | 0.79 | 0.54 | 0.64 | 0.92 | 0.93 | 1.31 |
| ACR | HJ-TrMBD | 0.99 | 0.98 | 0.97 | 0.95 | 0.94 | 0.92 |
| | HT-TrMBD | 0.94 | 0.91 | 0.89 | 0.89 | 0.89 | 0.88 |
| | HJ-LinMBD | 0.87 | 0.85 | 0.85 | 0.88 | 0.88 | 0.87 |
| | HT-LinEBLUP | 0.85 | 0.85 | 0.86 | 0.87 | 0.88 | 0.87 |
| AW | HJ-TrMBD | 1,592 | 22,688 | 140,452 | $52 \times 10^4$ | $35 \times 10^5$ | $44 \times 10^6$ |
| | HT-TrMBD | 219 | 4,414 | 34,105 | $14 \times 10^4$ | $11 \times 10^5$ | $15 \times 10^6$ |
| | HJ-LinMBD | 1,005 | 19,232 | 139,420 | $57 \times 10^4$ | $41 \times 10^5$ | $56 \times 10^6$ |
| | HT-LinEBLUP | 382 | 7,099 | 57,039 | $26 \times 10^4$ | $21 \times 10^5$ | $32 \times 10^6$ |

**Table 3**
**Average relative bias (ARB), average relative RMSE (ARRMSE), average coverage rate (ACR) and average interval width (AW) for model-based simulation Set B**

| Criterion | Estimator | $\gamma = -1.0$ | $\gamma = -0.5$ | $\gamma = 0.0$ | $\gamma = 0.5$ | $\gamma = 1.0$ |
|---|---|---|---|---|---|---|
| ARB,% | HT-TrMBD | 4.92 | 0.66 | 0.14 | -1.50 | -8.75 |
| | HJ-LinMBD | -0.21 | 0.04 | 0.12 | 0.16 | -0.85 |
| | HT-LinEBLUP | -0.19 | 0.04 | 0.13 | 0.17 | -0.77 |
| ARRMSE | HT-TrMBD | 0.38 | 0.35 | 0.33 | 0.37 | 0.41 |
| | HJ-LinMBD | 0.56 | 0.36 | 0.34 | 0.53 | 1.20 |
| | HT-LinEBLUP | 0.38 | 0.30 | 0.29 | 0.36 | 0.56 |
| ACR | HT-TrMBD | 0.94 | 0.92 | 0.92 | 0.91 | 0.87 |
| | HJ-LinMBD | 0.91 | 0.92 | 0.92 | 0.92 | 0.90 |
| | HT-LinEBLUP | 0.93 | 0.94 | 0.94 | 0.93 | 0.92 |
| AW | HT-TrMBD | 0.04 | 2.50 | 211 | 29,070 | $5 \times 10^6$ |
| | HJ-LinMBD | 0.06 | 2.70 | 214 | 38,660 | $13 \times 10^6$ |
| | HT-LinEBLUP | 0.05 | 2.60 | 214 | 33,442 | $10 \times 10^6$ |

### 6.2 The design-based simulation study

This study used the same population and samples as the simulation studies described in Chandra and Chambers (2005) and Chambers and Tzavidis (2006), which was based on data obtained from a sample of 1,652 farms that participated in the Australian Agricultural and Grazing Industries Survey (AAGIS). A realistic population of 81,982 farms was defined by sampling with replacement from the original sample of 1,652 farms with probabilities proportional to their sample weights, all of which were strictly greater than one. A total of 1,000 independent samples, each of size $n = 1{,}652$, were drawn from this fixed population by simple random sampling without replacement within strata defined by the 29 Australian agricultural regions represented in the AAGIS sample. These regions are the small areas of interest. Regional sample sizes were fixed to be the same as in this original sample, varying from a low of 6 to a high of 117, which allows an evaluation of the performance of the different estimation methods across a range of realistic small area sample sizes. Note that sampling fractions in these strata also varied disproportionately, ranging between 0.70

and 15.87 percent. The aim is to estimate average annual farm costs (TCC, measured in A$) in each region using farm size (hectares) as the auxiliary variable. The same mixed model specification as in Chandra and Chambers (2005) is used. This includes an interaction term (zone by size) in the fixed effects and a random slope specification for the area effect. In its linear form the model does not fit the AAGIS sample data terribly well. This fit is improved (albeit marginally) when a log-scale linear specification is used. Our results are summarized in Table 4.

### 6.3 Discussion of simulation results

The most striking feature of Table 2 is the extremely large values of the averages relative bias of HJ-TrMBD under model-based model-calibrated weighting. The two best performers with respect to relative bias are HT-TrMBD, which is based on the same weights as HJ-TrMBD, and TrEP. An investigation of the reason for the poor performance of HJ-TrMBD revealed that summing the model-based model-calibrated weights (17) within small areas produced extremely variable estimates of the small area population sizes, implying that these weights cannot be considered as 'multipurpose' – they function well when used with variables that are reasonably correlated with the variable that defines the fitted value model, but can fail with other, less well correlated, variables (*e.g.*, the indicator variable for small area inclusion). We further note that this problem does not arise with the 'standard' empirical EBLUP weights for the total (6), as HJ-LinMBD performs consistently for all six of the scenarios explored in Set A of the simulation study. From now on we therefore focus our discussion on the four estimators, HT-TrMBD, TrEP, HJ-LinMBD and HT-LinEBLUP.

Table 2 shows that the average relative biases and the average relative RMSEs for HT-TrMBD are consistently lower than those generated by HJ-LinMBD and HT-LinEBLUP. The average relative biases of HT-TrMBD and TrEP are comparable. However, the average relative RMSEs of HT-TrMBD are consistently smaller than the TrEP. Furthermore, average coverage rates and interval widths for HT-TrMBD are better than those generated by HJ-LinMBD and HT-LinEBLUP. In comparison, for the same order of relative bias, the relative RMSEs of HT-LinEBLUP is smaller than that of HJ-LinMBD, and, although both estimators generate very similar coverage rates, confidence intervals generated via HT-LinEBLUP tend to have smaller average widths than those generated via HJ-LinMBD.

The plots in Figure 1 display the region-specific performance measures generated by these four estimators for the Set A simulations. These show that the relative bias and the relative RMSE values generated by HT-TrMBD are smaller than corresponding values for HJ-LinMBD and HT-LinEBLUP in all regions. With almost identical values of relative biases, the HT-TrMBD has smaller values of relative RMSEs than corresponding values for TrEP in all regions. Further, the relative bias and the relative RMSE of HJ-LinMBD and HT-LinEBLUP increase as the non-linearity in the data increases (*i.e.*, as we move from parameter set 1 to parameter set 6). We also see that HT-TrMBD generates better coverage rates across all regions compared with the coverage rates generated by HT-LinEBLUP and HJ-LinMBD.

**Table 4**
**Average relative bias (ARB), average relative RMSE (ARRMSE) and average coverage rate (ACR) for design-based simulation using AAGIS data. Simulation standard errors of ARB and ARRMSE are shown in parentheses**

| Criterion | Estimator | Average of 29 regions | Average of 28 regions |
|---|---|---|---|
| ARB,% | HT-TrMBD | 1.96 (0.20) | 1.92 (0.11) |
| | HJ-LinMBD | -2.13 (0.15) | -2.21 (0.12) |
| | HT-LinEBLUP | 2.98 (0.18) | 3.36 (0.16) |
| | PseudoEBLUP | 4.01 (0.22) | 4.41 (0.20) |
| | JL | 1.89 (0.19) | 2.23 (0.17) |
| ARRMSE, % | HT-TrMBD | 21.93 (4.47) | 17.41 (1.18) |
| | HJ-LinMBD | 20.15 (3.80) | 16.91 (2.20) |
| | HT-LinEBLUP | 19.87 (1.78) | 19.30 (1.63) |
| | PseudoEBLUP | 22.42 (2.52) | 21.95 (2.46) |
| | JL | 20.97 (1.48) | 20.48 (1.31) |
| ACR | HT-TrMBD | 0.89 | 0.92 |
| | HJ-LinMBD | 0.93 | 0.95 |
| | HT-LinEBLUP | 0.85 | 0.85 |

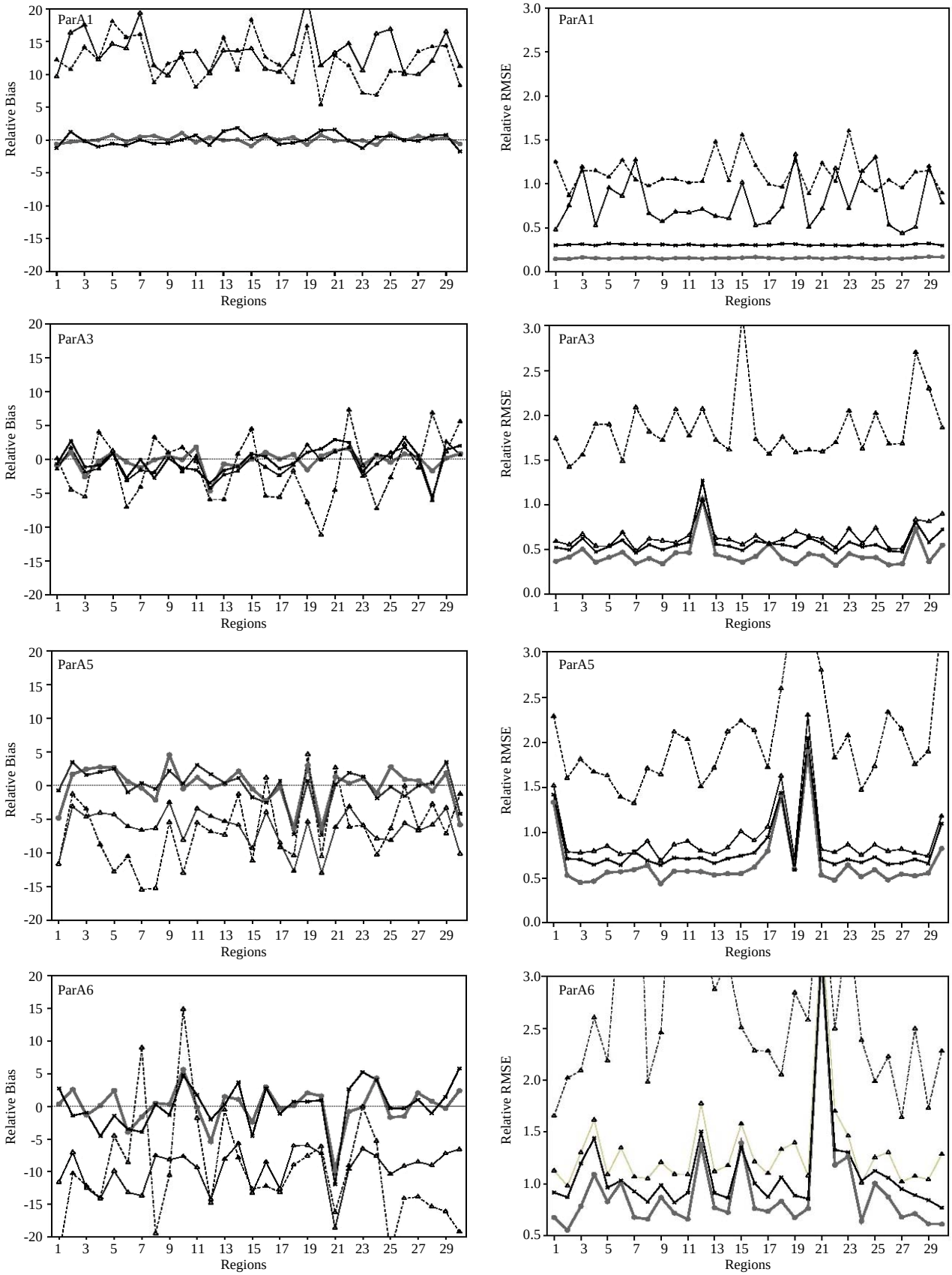


**Figure 1 Area specific results for HT-TrMBD (solid line, ●), TrEP (thick line, ×), HT-LinEBLUP (thin line Δ) and HJ-LinMBD (dashed line, Δ) under parameter sets 1 (ParA1), 3 (ParA3), 5 (ParA5) and 6 (ParA6). Left column is Relative Bias (%) and right column is Relative RMSE**

Overall, these results show that when the model for the underlying population is non-linear there can be significant gains from the use of HT-type MBD estimators for small area means based on the model-calibrated weights (17) compared with standard linear mixed model-based estimators like HJ-LinMBD and HT-LinEBLUP. They also show that the indirect estimator HT-LinEBLUP performs relatively better than the direct estimator HJ-LinMBD in these situations. The indirect predictor TrEP based on log-transformed model (9) performs well in terms of relative bias but is less efficient than the MBD estimator under the same model.

In Set B of the model-based simulations we investigated the robustness of model-based model-calibrated direct estimation to misspecification of the non-linear model. The results in Table 3 show that in this case the biases generated by HT-TrMBD increase as the actual non-linear model deviates more from the assumed non-linear model ($\gamma = 0.0$ in the table). However, these biases are offset by small variability, so in terms of average relative RMSE, HT-TrMBD still performs as well or better than HT-LinEBLUP and continues to dominate HJ-LinMBD. The biases generated by HJ-LinMBD and HT-LinEBLUP are of the same order, while the average relative RMSE of HT-LinEBLUP dominates that of HJ-LinMBD. Average coverage rates for HT-LinEBLUP are marginally better than those of HJ-LinMBD and HT-TrMBD, but the average widths of the confidence intervals underpinning these rates tended to be smallest for HT-TrMBD, followed by HT-LinEBLUP and then HJ-LinMBD. Overall, our model-based simulation results for Set B indicate that although MBD-based SAE with model-based model-calibrated weights is susceptible to model misspecification bias, the overall performance of this approach appears relatively unaffected by slight deviations from the assumed non-linear model.

In Table 4 and Figure 2 we present the average and region-specific performance measure generated by different SAE methods for AAGIS data respectively. These results show that the average relative bias of HT-TrMBD is smaller than that of both HT-LinEBLUP and HJ-LinMBD, while the average relative RMSE of HT-TrMBD is marginally larger than the corresponding values for HJ-LinMBD and HT-LinEBLUP. Inspection of Figure 2 shows that this result is essentially due to one region (21) in the original AAGIS sample that contained a massive outlier (TCC > A$30,000,000). This outlier was included in the simulation population (twice) and then selected (in one case, twice) in 37 of the 1000 simulation samples, leading to completely unrealistic estimates for region 21 being generated by HT-TrMBD and HJ-LinMBD. The right-hand column in Table 4 therefore shows the average performances of the different methods when this region is excluded. Here we see that now HT-TrMBD and HJ-LinMBD are essentially on a par, with both dominating HT-LinEBLUP. The fact that HT-TrMBD does not provide significant gains over HJ-LinMBD in this case reflects the fact that the raw-scale and log-scale linear mixed models used in these estimators both provide relatively poor fits to the AAGIS data.

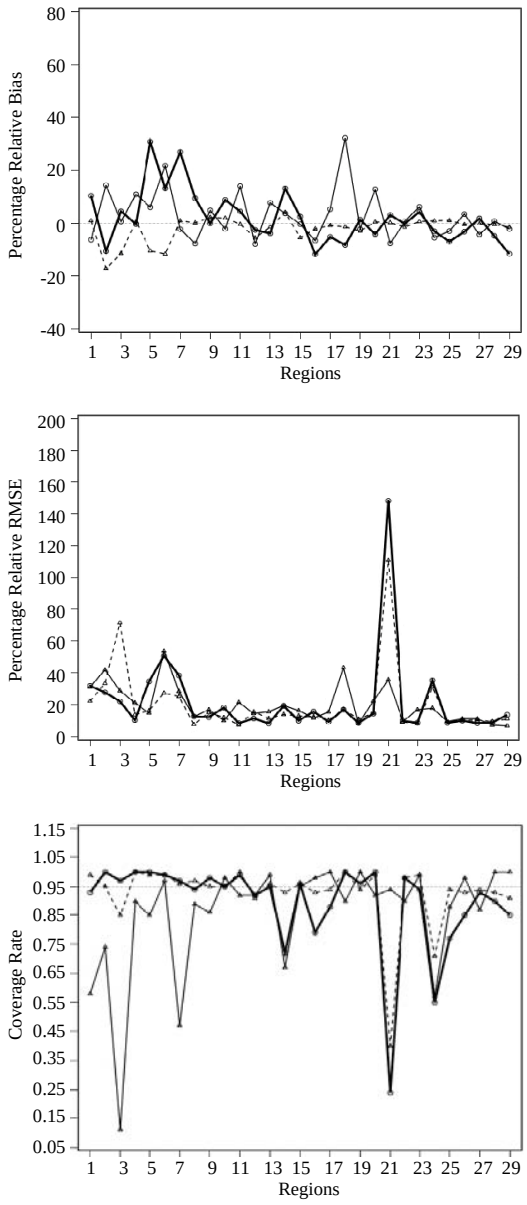


**Figure 2 Region-specific simulation results for HT-TrMBD (thick line, 0), HT-LinEBLUP (thin line Δ) and HJ-LinMBD (dashed line, Δ) in design-based simulations based on the AAGIS data. Plots show (in order from the top), RB (%), RRMSE (%) and CR. Regions are ordered in terms of increasing population size**

## 7. Conclusions and further research

The simulation results discussed in the previous section show that combining model-based model-calibrated weights with direct estimation can bring significant gains in SAE efficiency if the population data are clearly non-linear. As one would expect, these gains are less when the assumed non-linear model is misspecified. Although we do not provide the details, our conclusions were essentially unaffected when we carried out similar simulations using gamma distributed random effects.

Our main caveat concerning the use of the model-based model-calibrated weights (17) for SAE is their specificity. These weights do not appear to have the same 'multi-purpose' characteristics as standard EBLUP weights for the total based on linear mixed models. Further research is therefore required on how to build model-calibrated weights for SAE that are more 'general purpose'. It is to be expected that such weights would not be as efficient as the variable specific weights (17), but hopefully this will be more than offset by their increased utility. A further issue that is extremely important in practice is that positively skewed survey variables can also take zero (or even negative) values. For example, economic variables like debt and capital expenditure often take zero values, while variables defined as the difference of two non-negative quantities (*e.g.*, profit, which is the difference between income and expenditure) can be negative. Karlberg (2000b) uses a mixture model to characterise data that are a mix of zeros and strictly positive values. This type of model can be used in model-based model-calibrated weighting.

Finally, we note that using a transformation-based MBD approach where the usual linear model assumptions are only approximately valid (the situation considered in this paper) is not the only approach that has been suggested for this problem. Two alternative approaches in the literature are the pseudo-EBLUP (Rao 2003, section 7.2.7) and the model-assisted EB-type estimator of Jiang and Lahiri (2006). Recollect from (8) that the EBLUP is defined by replacing the unknown area $i$ mean $m_{iy}$ by an estimate of its expected value given the observed sample values of $Y$ in area $i$ and the area $i$ values of $\mathbf{X}$. Let $\pi_{ij}$ denote the sample inclusion probability of population unit $j$ in small area $i$. The pseudo-EBLUP is then defined by replacing $m_{iy}$ by an estimate of its expected value given the value of its design-consistent estimate

$$\hat{m}_{iy}^{\pi} = \left(\sum_{j\in s_i}\pi_{ij}^{-1}\right)^{-1}\sum_{j\in s_i}\pi_{ij}^{-1}y_{ij} = \sum_{j\in s_i}\tilde{w}_{ij}y_{ij} \quad (21)$$

and the area $i$ values of $\mathbf{X}$. That is, under (3) the pseudo-EBLUP of $m_{iy}$ is

$$\begin{aligned}\hat{m}_{iy}^{\text{psuedoEBLUP}} &= \hat{E}\{m_{iy} \mid \hat{m}_{iy}^{\pi}, \mathbf{x}_{is}, \mathbf{x}_{ir}\}\\ &= \overline{\mathbf{x}}_i'\hat{\beta}_{\tilde{w}} + (\overline{\mathbf{g}}_i'\hat{\Sigma}_{u\tilde{w}}\overline{\mathbf{g}}_{i\tilde{w}})\\ &\quad\left(\overline{\mathbf{g}}_{i\tilde{w}}'\hat{\Sigma}_{u\tilde{w}}\overline{\mathbf{g}}_{i\tilde{w}} + \hat{\sigma}_{e\tilde{w}}^2\sum\nolimits_{j\in s_i}\tilde{w}_{ij}^2\right)^{-1}(\hat{m}_{iy}^{\pi} - \overline{\mathbf{x}}_{i\tilde{w}}'\hat{\beta}_{\tilde{w}}) \quad (22)\end{aligned}$$

where $\hat{\beta}_{\tilde{w}}$, $\hat{\Sigma}_{u\tilde{w}}$ and $\hat{\sigma}_{e\tilde{w}}^2$ are pseudo-maximum likelihood estimates based on the weights $\tilde{w}_{ij}$ and $\overline{\mathbf{g}}_{i\tilde{w}}$ and $\overline{\mathbf{x}}_{i\tilde{w}}'$ are design-consistent estimates of $\overline{\mathbf{g}}_i$ and $\overline{\mathbf{x}}_i$ that are defined in exactly the same way as $\hat{m}_{iy}^{\pi}$ above. Under the same model the Jiang and Lahiri (2006) model-assisted EB-type approach leads to an estimator that is also defined by conditioning on the value of $\hat{m}_{iy}^{\pi}$,

$$\begin{aligned}\hat{m}_{iy}^{JL} &= \sum\nolimits_{j\in s_i}\tilde{w}_{ij}\hat{E}\{\hat{E}(y_{ij} \mid \mathbf{x}_{ij}, \mathbf{u}_i) \mid \hat{m}_{iy}^{\pi}, \mathbf{x}_i\}\\ &= \overline{\mathbf{x}}_{i\tilde{w}}'\hat{\beta} + \{\tilde{\mathbf{w}}_{is}'(\mathbf{g}_{is}\hat{\Sigma}_u\mathbf{g}_{is}' + \hat{\sigma}_e^2\mathbf{I}_{is})\tilde{\mathbf{w}}_{is}\}^{-1}\\ &\quad\{\tilde{\mathbf{w}}_{is}'\mathbf{g}_{is}\hat{\Sigma}_u\mathbf{g}'\tilde{\mathbf{w}}_{is}\}(\hat{m}_{iy}^{\pi} - \overline{\mathbf{x}}_{i\tilde{w}}'\hat{\beta}) \quad (23)\end{aligned}$$

where $\tilde{\mathbf{w}}_{is}$ is the vector of standardised sample weights $\tilde{w}_{ij}$ in area $i$. Note that in (23) we use optimal (*i.e.*, ML or REML) estimates for model parameters.

Both (22) and (23) are essentially motivated by the idea of estimating the area $i$ mean by its conditional expectation under (3) given the value of the usual design-consistent estimator (21) for this quantity. As such, they are indirect estimators like the HT-LinEBLUP. Under (3), neither will be as efficient as the HT-LinEBLUP, while if (9) rather than (3) holds, then both estimators rely on the design consistency of $\hat{m}_{iy}^{\pi}$ for robustness. Since relying on a large sample property of a small sample statistic seems rather optimistic, we prefer to tackle the model specification problem directly, replacing (3) by (9) and using the transformation-based MBD approach described in section 5. Values of average relative bias and average relative RMSE for the pseudo-EBLUP (22) and the Jiang and Lahiri estimator (23) are shown in Table 4. It is interesting to note that neither estimator appears to perform any better than the standard EBLUP in these design-based simulations, and all three are substantially out performed in terms of average relative RMSE by the two MBD-type estimators that were investtigated in this study. Clearly the results of a single (but reasonably realistic) simulation study should not be considered as anything more than indicative. However, they do provide some evidence that asymptotic design-based properties are no guarantee of small area estimation performance.

The indirect predictor (20) of the small area mean is obtained by using well known prediction-based ideas. Under log transformed models, there are alternative approaches to obtain better indirect predictor for small area mean. For example, Slud and Maiti (2006) described an

indirect predictor for the small area mean under an area level version of the log transformed model (9). Berg (2009, private communication) follows the Slud-Maiti approach to obtain a predictor for small area mean under a random intercepts specification of the unit level log transformed model (9). However, like the Slud-Maiti predictor, Berg's predictor ignores the bias correction necessary after back-transformation to the raw scale. The empirical properties of this predictor have yet to be examined.

## Acknowledgements

The first author gratefully acknowledges the financial support provided by a PhD scholarship from the U.K. Commonwealth Scholarship Commission. Constructive comments from Editor, Associate Editor and two referees are also gratefully acknowledged. They resulted in the revised version of the article representing a considerable improvement on the original.

## References

Bates, D.M., and Pinheiro, J.-C. (1998). Computational Methods for Multilevel Models. http://franz.stat.wisc.edu/pub/NLME/.

Carroll, R., and Ruppert, D. (1988). *Transformation and Weighting in Regression*. New York: Chapman and Hall.

Chambers, R., and Tzavidis, N. (2006). M-quantile models for small area estimation. *Biometrika*, 93, 255-268.

Chandra, H., and Chambers, R.L. (2005). Comparing EBLUP and C-EBLUP for small area estimation. *Statistics in Transition*, 7, 637-648.

Chandra, H., and Chambers, R. (2009). Multipurpose weighting for small area estimation. *Journal of Official Statistic*, 25(3), 379-395.

Chandra, H., Salvati, N. and Chambers, R. (2007) Small area estimation for spatially correlated populations. A comparison of direct and indirect model-based methods. *Statistics in Transition*, 8, 887-906.

Chen, G., and Chen, J. (1996). A transformation method for finite population sampling calibrated with empirical likelihood. *Survey Methodology*, 22, 139-146.

Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72, 320-338.

Hidiroglou, M.A., and Smith, P.A. (2005). Developing small area estimates for business surveys at the ONS. *Statistics in Transition*, 7, 527-539.

Jiang, J., and Lahiri, P. (2006). Estimation of finite population domain means: A model-assisted empirical best prediction approach. *Journal of the American Statistical Association*, 101, 301-311.

Karlberg, F. (2000a). Population total prediction under a lognormal superpopulation model. *Metron*, LVIII, 53-80.

Karlberg, F. (2000b). Survey estimation for highly skewed populations in the presence of zeroes. *Journal of Official Statistics*, 16, 229-241.

Longford, N.T. (2007). On standard errors of model-based small-area estimators. *Survey Methodology*, 33, 69-79.

McCulloch, C.E., and Searle, S.R. (2001). *Generalized, Linear and Mixed Models*. New York: John Wiley & Sons, Inc.

Prasad, N.G.N., and Rao, J.N.K. (1990). The estimation of the mean squared error of small area estimators. *Journal of the American Statistical Association*, 85, 163-171.

Rao, J.N.K. (2003). *Small Area Estimation*. New York: John Wiley & Sons, Inc.

Royall, R.M. (1976). The linear least-squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, 71, 657-664.

Royall, R.M., and Cumberland, W.G. (1978). Variance estimation in finite population sampling. *Journal of the American Statistical Association*, 73, 351-358.

Slud, E. V., and Maiti, T. (2006). Mean squared error estimation in transformed Fay–Herriot models. *Journal of the Royal Statistical Society*, Series B, 68(2), 239-257.

Tzavidis, N., Salvati, N., Pratesi, M. and Chambers, R. (2008). M-quantile models with application to poverty mapping. *Statistical Methods And Applications*, 17, 393-411.

Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite Population Sampling and Inference*. New York: John Wiley & Sons, Inc.

Wu, C., and Sitter, R.R. (2001). A model calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical* Association, 96, 185-193.


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

Survey Methodology, June 2011
Vol. 37, No. 1, pp. 53-65
Statistics Canada, Catalogue No. 12-001-X



# The construction of stratified designs in R with the package *stratification*

**Sophie Baillargeon and Louis-Paul Rivest** [1]

Abstract

This paper introduces a R-package for the stratification of a survey population using a univariate stratification variable $X$ and for the calculation of stratum sample sizes. Non iterative methods such as the cumulative root frequency method and the geometric stratum boundaries are implemented. Optimal designs, with stratum boundaries that minimize either the CV of the simple expansion estimator for a fixed sample size $n$ or the $n$ value for a fixed CV can be constructed. Two iterative algorithms are available to find the optimal stratum boundaries. The design can feature a user defined certainty stratum where all the units are sampled. Take-all and take-none strata can be included in the stratified design as they might lead to smaller sample sizes. The sample size calculations are based on the anticipated moments of the survey variable $Y$, given the stratification variable $X$. The package handles conditional distributions of $Y$ given $X$ that are either a heteroscedastic linear model, or a log-linear model. Stratum specific non-response can be accounted for in the design construction and in the sample size calculations.

Key Words: Linear models; Log-linear models; Optimal stratification; Survey sampling; Take-all stratum; Take-none stratum.

## 1. Introduction

The establishment of strata and the planning of a stratified design have been important topics in survey sampling, since the pioneering contributions of Dalenius more than sixty years ago. This work is concerned with univariate stratification where the strata are constructed using a positive stratification variable $X$ known for all the units of the population. $X$ is assumed to be related to the survey variable $Y$. Stratum $h$ contains all the units with an $X$-value in the interval $[b_{h-1}, b_h)$ for $h = 1, \ldots, L$ such that $b_0 = \min X$ and $b_L = \max X + 1$, where $\min X$ and $\max X$ are respectively the minimum and the maximum values of the stratification variable.

The determination of optimal stratum boundaries has a long history, see chapter 5A of Cochran (1977). The cumulative root frequency method (cum $\sqrt{f}$) of Dalenius and Hodges (1959) provides an approximate solution to this problem. Instances where $X$ has a skewed distribution are frequent in business surveys and have been given a special emphasis. Gunning and Horgan (2004) proposed a geometric stratification method and Hidiroglou (1986) argued that the large units should be put in a take-all stratum. Rather than relying on an approximate method for constructing the strata, Lavallée and Hidiroglou (1988) suggested an iterative algorithm that gives the optimal boundaries for a particular $X$ variable. Their algorithm sometimes fails to converge (Detlefsen and Veum 1991) and Slanta and Krenzke (1996) have shown that in some cases the optimal boundaries are not uniquely defined. Alternative methods, such as the search algorithm of Kozak (2004), have been proposed to alleviate some of these difficulties. The assumption that the survey variable $Y$ is the same as the stratification variable $X$ is not realistic when calculating sample sizes and several authors, including Dayal (1985) and Sigman and Monsour (1995), proposed to allocate the sample to the strata on the basis of the *anticipated* moments of $Y$ knowing that $X$ is in $[b_{h-1}, b_h)$. Sweet and Sigman (1995) and Rivest (1999, 2002) suggested using these anticipated moments in the stratification algorithm of Lavallée and Hidiroglou (1988). Recently, Baillargeon and Rivest (2009) showed that putting the small units in a take-none stratum, which is not sampled, might reduce the sample size needed to reach a predetermined precision level.

This article introduces the R-package *stratification* that implements most of the methods presented above. It provides a friendly computer environment to build stratified designs and to evaluate their performance on some real populations. This package is presented by revisiting examples in the stratification literature selected to illustrate its important features. The four functions of *stratification* with the prefix `strata` construct stratified sampling designs. These functions are `strata.cumrootf`, `strata.geo`, `strata.LH`, and `strata.bh`. The first two implement the simple cum $\sqrt{f}$ and geometric stratification methods. The function `strata.LH` derives optimal stratified sampling plans using iterative algorithms while the last function handles user defined stratum boundaries. These four functions construct strata, determine stratum sample sizes and calculate the precision of the simple expansion estimator $\bar{y}_s$ of $\bar{Y}$, the population mean of some survey variable $Y$ related to the stratification variable $X$.

1. Sophie Baillargeon, Département de mathématiques et de statistique, 1045, avenue de la médecine, Université Laval, Québec, (Qc) Canada G1V 0A6. E-mail: sophie.baillargeon@mat.ulaval.ca; Louis-Paul Rivest, Département de mathématiques et de statistique, 1045, avenue de la médecine, Université Laval, Québec, (Qc) Canada G1V 0A6. E-mail: louis-paul.rivest@mat.ulaval.ca.

The four `strata`-functions use Hidiroglou and Srinath's (1993) rule to allocate the $n$ units in the sample to the strata. The stratum sample sizes are proportional to $N_h^{2q_1}$ $\overline{Y}_h^{2q_2}$ $S_{yh}^{2q_3}$, where $N_h$ is the size of stratum $h$, and $\overline{Y}_h$ and $S_{yh}^2$ are the anticipated mean and variance of $Y$ in stratum $h$. In the `strata`-functions, an allocation rule is specified by the argument `alloc` that contains the exponents $(q_1, q_2, q_3)$; Neyman's allocation corresponds to `alloc=c(1/2,0,1/2)`. A `strata`-function takes as an input the population vector of the stratification variable $X$, the number of strata `Ls`, and a total sample size $n$ or a target CV for the simple expansion estimator $\overline{y}_s$. Its output is an `R`-object of class "strata" that defines a stratified design. It contains a set of strata determined by their upper boundaries $\{b_h\}$ and stratum population and sample sizes, $N_h$ and $n_h$. There is a fifth function in *stratification* called `var.strata` that takes as an input an `R`-object of class strata and a population vector of a survey variable $Y$ and returns the variance of $\overline{y}_s$ for the input variable $Y$ and the input stratified design.

The text contains `R` instructions to be typed in an `R` command window; these lines start with >. It also presents outputs printed in an R command window. A special typeface allows an easy identification of these `R` instructions and print-outs in the text. The appendix contains a summary table that lists all the possible arguments of the five *stratification* functions. When using this package, the `R`-instruction `help(stratification)` calls a clickable help file that provides detailed information on the package and examples that can be pasted in a command window.

## 2. Basic stratification methods

This section discusses two elementary stratification methods, the cumulative root frequency method of Dalenius and Hodges (1959) and the geometric method of Gunning and Horgan (2004). These two methods are exact; they do not rely on an iterative algorithm. Throughout this section $Y = X$, so that the variance of $\overline{y}_s$ is evaluated using the values of the stratification variable $X$. Using the same variable to stratify a population and to evaluate the precision of survey estimates might underestimate their variances. The calculation of variances when $Y \neq X$ is considered in Section 4.

### 2.1 Cumulative root frequency method

This stratification algorithm, presented in chapter 5A of Cochran (1977), is implemented by the function `strata.cumrootf`. Its arguments are `x`, the population vector of the stratification variable, `nclass` the number of bins of equal size for the `x`-variable, a target `CV` for $\overline{y}_s$ or a predetermined sample size `n`, the number of strata `Ls`, and an allocation rule `alloc`. This algorithm pools the `nclass` bins into `Ls` strata in such a way that the sums of the square roots of the bin frequencies are approximately equal for the `Ls` strata.

As an illustration, consider the proportion of industrial loans of $N$ = 13,435 banks used in Cochran (1961). We stratify this population and evaluate the sample size needed for $\overline{y}_s$ to have a CV of 5% when Neyman allocation is used. The following `R`-code creates the vector of the stratification variable `loans` from Table 2 of McEvoy (1956). The function `strata.cumrootf` is then applied to the loans variable. Following Table 2 of Cochran (1961), `nclass` is set to 20 so that the strata will be created using 20 bins and `Ls=3` strata will be constructed. The output is placed in `cum`, an `R`-object of class strata. Typing `cum` or `print(cum)` in the `R` command window prints details of the sampling plan. The input arguments, either the default or as specified by the user, appear first. Then stratum information is provided such as boundaries, sizes $N_h$ and sample sizes $n_h$. The third part of the print-out provides information about the sampling properties of $\overline{y}_s$.

```
> values <- c(seq(0.5, 9.5, 1), seq(12.5, 97.5, 5))
> nrep <- c(1985, 261, 339, 405, 474, 478, 506, 569, 464, 499,
  2157, 1581, 1142, 746, 512, 376, 265, 207, 126, 107, 82, 50,
  39, 25, 16, 19, 2, 3)
> loans <- rep(values, nrep)
> cum <- strata.cumrootf(x = loans, nclass = 20, CV = 0.05,
  Ls = 3, alloc = c(0.5, 0, 0.5))
> cum

Given arguments:
x = loans
nclass = 20, CV = 0.05, Ls = 3
allocation : q1 = 0.5, q2 = 0, q3 = 0.5
model = none

Strata information:
           rh |   bh anticip.Mean anticip.var    Nh   nh    fh
Stratum 1   1 | 10.2         4.12       10.46  5980   14  0.00
Stratum 2   1 | 29.6        17.92       27.74  5626   20  0.00
Stratum 3   1 | 98.5        44.47      165.83  1829   16  0.01
Total                                    13435   50  0.00

Total sample size: 50
Anticipated population mean: 15.39408
Anticipated CV: 0.0494897
```

In the `Given arguments, model=none` means that the sampling properties of $\overline{y}_s$, presented at the end of the print-out, are evaluated at $Y = X$, that is for the `loans` variable. Its mean is 15.39408 and the anticipated `CV` of 0.0494897 is that of the estimator $\overline{y}_s$ of the mean of the variable `loans` obtained with this sampling design. The stratum boundaries given in this output are (10.2, 29.6, 98.5), they are equal to those appearing at the bottom of page 349 of Cochran (1961), once the rounding used for creating the vector `loans` is accounted for. In the `Strata Information`, $r_h$ refers to the stratum response rates that are discussed in Section 5.1. The `R`-object `cum` contains several elements that are listed by the command `names(cum)`.

```
> names(cum)
 [1] "Nh"           "nh"   "n"          "nh.nonint" "certain.info"
 [6] "opti.criteria" "bh"   "meanh"      "varh"      "mean"
[11] "stderr"       "CV"   "stratumID"  "nclassh"   "takeall"
[16] "call"         "date" "args"
```

An element in the cum strata object can be printed by typing cum$ followed by the name of the object. For instance the cum$stratumID prints the stratum of each unit in the population. The variable cum$nclassh is specific to the strata.cumrootf function; it gives how the nclass=20 original bins have been pooled into three strata;

```
> cum$nclassh
[1] 2 4 14
```

Thus, in this stratification, strata 1, 2 and 3 contain respectively 2, 4 and 14 of the nclass=20 original bins.

### 2.2 Geometric method

The geometric stratification method has been introduced by Gunning and Horgan (2004). It sets the stratum boundaries to $b_h = \min X \times (\max X / \min X)^{h/L}$, for $h = 1, ..., L-1$. Once the boundaries $b_h$ are determined, the stratum sample size calculations are the same as those carried out in strata.cumrootf.

As an illustration we stratify the four populations presented in Gunning et Horgan (2004), Debtors, USbanks, UScities, and UScolleges, into Ls=5 strata. The last three populations were considered in Cochran's (1961) investigations. These four populations are stored in *stratification*; the command data(Debtors) calls the first one. Rather than specifying a target CV we set the sample size to $n$ = 100 following Gunning and Horgan (2004). The following commands create the R-object pop1 that contains the stratified design for the Debtors population.

```
> data(Debtors)
> pop1 <- strata.geo(x = Debtors, n = 100, Ls = 5,
  alloc = c(0.5, 0, 0.5))
```

Table 1 summarizes the geometric stratified designs for the four study populations. It reproduces Table 4 of Gunning and Horgan (2004) partially. There are however some minor differences caused by different rounding strategies. More details about *stratification* rounding methods are available in the help file.

**Table 1**
**Stratified designs for four populations with $n$ = 100**

| Population | CV | | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|---|
| Debtors | 0.0359 | $b_h$ | 148.28 | 549.67 | 2,037.60 | 7,553.33 | |
| | | $N_h$ | 1,054 | 1,267 | 732 | 265 | 51 |
| | | $n_h$ | 3 | 14 | 27 | 33 | 23 |
| UScities | 0.0145 | $b_h$ | 18.17 | 33.01 | 59.98 | 108.98 | |
| | | $N_h$ | 364 | 418 | 130 | 87 | 39 |
| | | $n_h$ | 18 | 28 | 17 | 20 | 17 |
| UScolleges | 0.0183 | $b_h$ | 434.00 | 941.76 | 2,043.61 | 4,434.60 | |
| | | $N_h$ | 94 | 255 | 198 | 74 | 56 |
| | | $n_h$ | 3 | 15 | 27 | 20 | 35 |
| USbanks | 0.0107 | $b_h$ | 118.59 | 200.92 | 340.39 | 576.68 | |
| | | $N_h$ | 114 | 116 | 64 | 39 | 24 |
| | | $n_h$ | 13 | 20 | 25 | 18 | 24 |

### 2.3 Take-all stratum

In Table 1, the fifth stratum for the USbanks population is a take-all stratum since $n_5 = N_5 = 24$. Under Neymann allocation, the fifth stratum gets a sample size $n_5$ larger than the stratum size $N_5$. Then strata.geo automatically identifies this stratum as a take-all stratum and allocates the $n - N_5$ units for the first four strata using Neyman allocation. This adjustment is important to have a sample size of $n = 100$ as specified in the strata.geo arguments.

To illustrate this point, we use the function strata.bh to make an allocation without a take-all stratum adjustment. This function allocates the sample and calculates the precision of $\bar{y}_s$ for a predetermined set of stratum boundaries. By setting takeall.adjust=FALSE, Neyman allocation is used in the five strata and since $n_5 > N_5$ one has $n_5 = N_5$. The following R-code gets the geometric stratum boundaries $\{b_h\}$ in the strata object adjust; it then uses the strata.bh function with the geometric stratum boundaries to get the sampling design without adjusting for a take-all stratum five in the noadjust strata object.

```
> data(USbanks)
> adjust <- strata.geo(x = USbanks, n = 100, Ls = 5,
  alloc = c(0.5, 0, 0.5))
> noadjust <- strata.bh(x = USbanks, bh = adjust$bh,
  n = 100, Ls = 5, alloc = c(0.5, 0, 0.5), takeall = 0,
  takeall.adjust = FALSE)
```

The two designs are presented in Table 2. Failing to include a take-all stratum yields a sample size of $n$ = 99, smaller than the target $n$ = 100. In this case, the unrounded sample size for stratum 5 is noadjust$nh.noint[5]= 25.40 for $N_5$ = 24 units. Note that when $n$ is large or when the target CV is small, it is possible to get several take-all strata.

**Table 2**
**Stratified designs obtained with and without an automatic adjustment for a take-all stratum**

| | $n$ | | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|---|
| | | $b_h$ | 118.59 | 200.92 | 340.39 | 576.68 | |
| | | $N_h$ | 114 | 116 | 64 | 39 | 24 |
| adjust | 100 | $n_h$ | 13 | 20 | 25 | 18 | 24 |
| noadjust | 99 | $n_h$ | 13 | 20 | 24 | 18 | 24 |

### 2.4 Adding a take-all stratum

We now consider the data base on $N$ = 284 Swedish municipalities given in the appendix of Särndal, Swensson and Wretman (1992). The following instructions use the geometric method to stratify this population in Ls=5 strata using the variable REV84, the 1984 real estate values. The power allocation with exponent 0.7 and alloc=c(0.35,0.35,0) is used. The R-object of class strata geo contains the stratified design. The command plot(geo) produces the plot presented in Figure 1. It provides a histogram of the

stratification variable with the stratum boundaries and a summary table for the stratified design.

```
> data(Sweden)
> geo <- strata.geo(x = Sweden$REV84, CV = 0.05, Ls = 5,
  alloc = c(0.35, 0.35, 0))
```

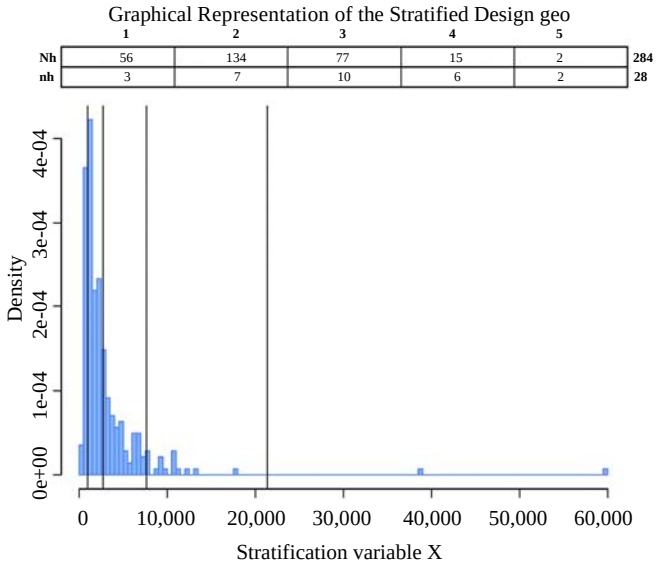


**Figure 1 Plot of the R-object geo**

Figure 1 shows that the geometric stratification method puts two of the three extreme REV84 values in a take-all stratum. The following Rcode creates cum a stratified design for this population using the cum $\sqrt{f}$ method. The application of this stratification method is awkward since the bins have length {max(REV84)−min(REV84)}/50= 1191. Considering Figure 1 most of the bins have a null frequency; indeed stratum 5 comprises 43 of the 50 bins. This design does not have a take-all stratum. To calculate the sample sizes obtained by requesting a take-all stratum one can use the function strata.bh, with the cum $\sqrt{f}$ boundaries stored in cum$bh, with the command takeall=1. This gives the third sampling plan in Table 3. The fourth sampling plan of Table 3 cum3 is created by setting the sample size in stratum 5 of the cum $\sqrt{f}$ design equal to its population size with the command cum3$nh[5]<−cum3$Nh[5]. The variance of the estimate $\bar{y}_s$ for the variable REV84 using this fourth sampling design is calculated using var.strata.

```
> cum <- strata.cumrootf(x = Sweden$REV84, nclass = 50,
  CV = 0.05, Ls = 5, alloc = c(0.35, 0.35, 0))
> cum2 <- strata.bh(x = Sweden$REV84, bh = cum$bh, CV = 0.05,
  Ls = 5, takeall = 1, alloc = c(0.35, 0.35, 0))
> cum3 <- cum
> cum3$nh[5] <- cum3$Nh[5]
> cum3.var <- var.strata(cum3, y = Sweden$REV84)
```

**Table 3**
**Four stratified designs for the population of Swedish municipalities**

| Method | | 1 | 2 | 3 | 4 | 5 | *n* | CV |
|---|---|---|---|---|---|---|---|---|
| geometric | $N_h$ | 56 | 134 | 77 | 15 | 2 | | |
| | $n_h$ | 3 | 7 | 10 | 6 | 2 | 28 | 4.83 |
| cum $\sqrt{f}$ | $N_h$ | 120 | 70 | 52 | 27 | 15 | | |
| | $n_h$ | 7 | 7 | 9 | 8 | 10 | 41 | 4.87 |
| | $n_h^{modif1}$ | 2 | 2 | 3 | 2 | 15 | 24 | 4.44 |
| | $n_h^{modif2}$ | 7 | 7 | 9 | 8 | 15 | 46 | 2.29 |

Table 3 highlights that the sampling fraction in the fifth stratum drives the value of *n*. The cum $\sqrt{f}$ design appears to be less efficient than the geometric design since it sampling fraction in stratum 5 is 10/15 = 67%. Requesting a take-all stratum gives a value of *n* comparable to that obtained with the geometric design. The REV84 population has three outliers that were identified in Table 1. The geometric and cum $\sqrt{f}$ stratification methods depend heavily on the maximum $X$-value; therefore before applying these techniques it might be wise to put the three outliers aside. This is considered in the next section.

The simple *ad hoc* method to arbitrarily change the stratum sample sizes presented in this section can be applied in several situations. For instance, when some strata have samples of size 1, they can be increased to 2 in order to have an unbiased variance estimator.

## 2.5 Certainty stratum

In a stratified design it might be useful to constrain some units to be sampled, before constructing the strata. The argument certain available in the four strata-function makes this possible. As an example we revisit the comparison of the cum $\sqrt{f}$ and the geometric sampling designs presented in Table 3. The three large municipalities highlighted in Figure 1 are put in a certainty stratum, and the $N = 281$ remaining municipalities are stratified into Ls=4 strata using the two stratification methods. The R-code for constructing these two designs is given below. The command x=sort(Sweden$REV84) orders the municipalities by increasing *REV*84; thus the three large municipalities are entries 282, 283 and 284 of the sorted vector. The two R objects of class strata, geo_cer and cum_cer, each contain an element certain.info that provides information on the certainty stratum.

```
> geo_cer <- strata.geo(x = sort(Sweden$REV84), CV = 0.05,
  Ls = 4, alloc = c(0.35, 0.35, 0), certain = 282:284)
> cum_cer <- strata.cumrootf(x = sort(Sweden$REV84),
  nclass = 50, CV = 0.05, Ls = 4, alloc = c(0.35, 0.35, 0),
  certain = 282:284)
> cum_cer$certain.info

   Nc      meanc
 3.00   38923.67
```

In Table 4, the cum $\sqrt{f}$ design is more efficient that the geometric design. Putting the three large municipalities in a certainty stratum is helpful since the sample sizes in Table 4 are smaller than those of Table 3. The argument `certain` can force any set of units in the sample. It can be used to include units that are extreme for a secondary variable, different from the stratification variable, or that have a history of high volatility.

**Table 4**
**Two stratified designs for the Swedish municipalities constructed with a certainty stratum**

| Method | | 1 | 2 | 3 | 4 | 5 | *n* | CV |
|---|---|---|---|---|---|---|---|---|
| geometric | $N_h$ | 42 | 116 | 88 | 35 | 3 | | |
| | $n_h$ | 2 | 5 | 7 | 7 | 3 | 24 | 4.71 |
| cum $\sqrt{f}$ | $N_h$ | 127 | 79 | 46 | 29 | 3 | | |
| | $n_h$ | 3 | 4 | 4 | 5 | 3 | 19 | 4.72 |

## 3. Optimization method

The stratification methods introduced in Section 2 do not always give an optimal stratified design, that minimizes the sample size $n$ needed to reach the target CV (or minimizes the CV for a fixed $n$). This section introduces the function `strata.LH` that allows the determination of optimal designs. The name `LH` stands for Lavallée and Hidiroglou (1988) who pioneered the construction of optimal stratified designs for real life survey populations. In a stratified design with a take-all stratum, the variance of the simple expansion estimator is given by

$$\mathrm{Var}(\bar{y}_s) = \sum_{h=1}^{L-1}\left(\frac{N_h}{N}\right)^2\left(\frac{1}{(n-N_L)a_h} - \frac{1}{N_h}\right)S_{yh}^2,$$

where $\{a_h\}$ is the allocation rule for setting stratum sample sizes. The $n$ that ensures a CV of $c$ is given by

$$n = N_L + \frac{\sum_{h=1}^{L-1} N_h^2 S_{yh}^2 \,/\, (a_h N^2)}{\bar{Y}^2 c^2 + \sum_{h=1}^{L-1} N_h S_{yh}^2 \,/\, N^2}. \qquad (1)$$

In this expression one can write $n = n(b_1,\ldots,b_L)$ to highlight that the value of $n$ depends on the stratum boundaries. The `strata.LH` function tries to find the optimal boundaries $b_h$ that minimize $n(b_1, \ldots, b_{L-1})$. Two minimization algorithms are available, either Sethi's (1963) algorithm as implemented by Lavallée and Hidiroglou (1988) with `algo="Sethi"` or Kozak's (2004) random search algorithm with `algo="Kozak"`. The latter is the default option. This section assumes $Y = X$; it does not distinguish the stratification from the survey variable.

### 3.1 Sethi (1963) example with the normal distribution

A classical problem is to determine the optimal boundaries for $L$ strata in an infinite population from a known distribution. For instance, Sethi (1963) derived the optimal bounds for the normal and the $\chi^2_{30}$ distributions. To obtain approximate solutions, one can run the `strata.LH` function on a large Monte Carlo population simulated from the known distribution, without requesting a take-all stratum. In (1), one has $N_h / N^2 \approx 0$ and the optimal boundaries are the same for any target CV $c$.

The following R-code simulates populations of size $10^5$ from the $N(10, 1)$ and the $\chi^2_{30}$ distributions. Observe that *stratification* requires the stratification variable to be non negative, so that it would not work on standard normal deviates. By subtracting 10 from the $N(10, 1)$ boundaries, we get the ones for a $N(0, 1)$. The calculations are done with the `strata.LH` function with the argument `algo="Sethi"` and with `takeall=0`, so that a take-all stratum is not requested.

```
> z <- rnorm(100000, 10)
> zl5 <- strata.LH(x = z, CV = 0.001, Ls = 5,
  alloc = c(0.5, 0, 0.5), takeall = 0, algo = "Sethi")
> zl5$bh - 10

[1] -1.1247340 -0.3480829 0.3297044 1.0979017

> x30 <- rchisq(100000, 30)
> xl5 <- strata.LH(x = x30, CV = 0.01, Ls = 5,
  alloc = c(0.5, 0, 0.5), takeall = 0, algo = "Sethi")
> xl5$bh

[1] 22.82148 28.12303 33.38642 40.20165
```

In Table 5, the agreement between the true bounds reported in Table 8 of Sethi (1963) and the Monte Carlo bounds is quite good. This approach could be used to calculate the optimal stratum boundaries for an arbitrary distribution, see for instance Khan, Nand, and Ahmad (2008).

**Table 5**
**Comparison of Sethi's (1963) optimal stratum boundaries and of the approximate boundaries obtained with *stratification***

| | | ***stratification*'s results** | | | | **Sethi's results** | | | |
|---|---|---|---|---|---|---|---|---|---|
| | ***L*** | **1** | **2** | **3** | **4** | **1** | **2** | **3** | **4** |
| | 2 | -0.007 | | | | 0.00 | | | |
| $b_h$ | 3 | -0.531 | 0.567 | | | -0.55 | 0.55 | | |
| $N(0,1)$ | 4 | -0.883 | -0.008 | 0.864 | | -0.88 | 0.00 | 0.88 | |
| | 5 | -1.125 | -0.348 | 0.330 | 1.098 | -1.11 | -0.34 | 0.34 | 1.11 |
| | 2 | 30.674 | | | | 30.6 | | | |
| $b_h$ | 3 | 26.535 | 35.141 | | | 26.0 | 35.0 | | |
| $\chi^2_{30}$ | 4 | 24.340 | 30.733 | 38.179 | | 24.0 | 30.6 | 38.0 | |
| | 5 | 22.821 | 28.123 | 33.386 | 40.202 | 22.0 | 28.0 | 33.0 | 40.0 |

### 3.2 Gunning and Horgan (2004) example

In their original proposal, Lavallée and Hidiroglou (1988) always had a take-all stratum for a skewed survey

variable. To show that this was not always mandatory, Gunning and Horgan (2004) derived the optimal stratified designs featuring a take-all stratum for the four populations considered in Table 1. The findings of their Table 7 (with slight corrections due to rounding errors) is reproduced in Table 6. Comparing Tables 1 and 6, one sees that the optimal designs featuring a take-all stratum have $n$-values larger than 100 for three populations out of four. The optimal design is superior to the geometric design only for the Debtors population. The R-code to run Sethi's algorithm on the Debtors population is given below.

```
> pop1LH <- strata.LH(x = Debtors, CV = 0.0359, Ls = 5,
  alloc = c(0.5, 0, 0.5), takeall = 1, algo = "Sethi")
```

In Table 6, one would expect the optimal designs obtained through an iterative algorithm to have a smaller sample size than the *ad hoc* geometric designs. This fails to occur for three populations. This might be caused by a failure of Sethi's algorithm to find the true minimum value for *n*. To check this, we reran the programs to produce Table 6 with the argument `algo="Kozak"`. The sample sizes $n$ are given in the second column of Table 7. Kozak's algorithm finds a smaller $n$-value that Sethi's for three of the four populations. This highlights the weakness of Sethi's algorithm for real populations. The second column of Table 7 has $n$ values larger than 100 for two of the four populations. In these cases, the geometric design might be better because a take-all stratum is not required. To check this we reran Kozak's algorithm wihout a take-all stratum, *i.e.*, with `takeall=0`. The results are reported in the third column of Table 7. For the Debtors and the UScolleges populations, taking away the take-all stratum reduces the sample size $n$. Still, for the UScities population, Kozak's algorithm does worst than the geometric design. It failed to find the true minimum value of $n$ with the default arguments that control its random search. To better understand the results of Table 7, we now present in more details the selection of initial stratum boundaries in `strata.LH` and the parameters that control the random search with `algo="Kozak"`.

**Table 6**
**Optimal stratified designs featuring a take-all stratum obtained with Sethi's algorithm for the 4 populations of Table 1**

| Population | $n$ | CV | | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|---|---|
| Debtors | 93 | 0.0359 | $b_h$ | 349.33 | 1,190.16 | 3,482.98 | 10,322.50 | |
| | | | $N_h$ | 1,856 | 991 | 350 | 146 | 26 |
| | | | $n_h$ | 13 | 17 | 17 | 20 | 26 |
| UScities | 137 | 0.0145 | $b_h$ | 14.72 | 21.62 | 35.59 | 80.47 | |
| | | | $N_h$ | 189 | 270 | 336 | 164 | 79 |
| | | | $n_h$ | 4 | 8 | 16 | 30 | 79 |
| UScolleges | 107 | 0.0183 | $b_h$ | 512.32 | 869.76 | 1,577.23 | 3,668.85 | |
| | | | $N_h$ | 133 | 180 | 185 | 110 | 69 |
| | | | $n_h$ | 4 | 6 | 10 | 18 | 69 |
| USbanks | 104 | 0.0107 | $b_h$ | 99.37 | 129.60 | 181.94 | 317.36 | |
| | | | $N_h$ | 70 | 66 | 82 | 65 | 74 |
| | | | $n_h$ | 4 | 4 | 7 | 15 | 74 |

**Table 7**
**Sample size $n$ for three optimal designs and four populations**

| Population | `algo=Sethi` | `algo=Kozak` | `algo=Kozak` |
|---|---|---|---|
| | `takeall=1` | `takeall=1` | `takeall=0` |
| Debtors | 93 | 92 | 82 |
| UScities | 137 | 114 | 123 |
| UScolleges | 107 | 107 | 95 |
| USbanks | 104 | 88 | 88 |

### 3.3 Customization of the algorithms

The default initial stratum boundaries for the two iterative algorithms are the arithmetic starting point of Gunning and Horgan (2007), with $b_h = \min X + (\max X - \min X) \times h/L$, for $h = 1,\ldots,L-1$. In Table 7, this choice is questionable and the geometric stratum boundaries would have been closer to the true optimal boundaries. In `strata.LH`, the argument `initbh=` allows to specify a vector of $L-1$ initial boundary values. The maximum number of iterations can be changed with the `maxiter` element of the `algo.control` argument.

Kozak's algorithm was first proposed in Kozak (2004), see also Kozak and Verma (2006). It uses a random search that selects the $L-1$ stratum boundaries among the sorted values of $X$, with the duplicates discarded. At one iteration, it randomly picks a number $d$ in the set `{-maxstep,-maxstep+1,…,maxstep}` and one of the $L-1$ boundaries. Then it moves the selected boundary by $d$ positions in the vector of sorted $X$-values. If (1) is smaller with the new boundary it is kept, otherwise it is discarded and the boundaries are left unchanged at this iteration. The algorithm stops when the boundaries have not been changed for `maxstill` consecutive iterations. The default values are `maxstep=3` and `maxstill=100`. Two consecutive runs of Kozak's algorithm might lead to different designs because of the random nature of this algorithm. The `strata.LH` runs the algorithm `rep` times and the information for each run is contained in the `rep.detail` element of `R`-objects of class strata; the default value is `rep=3`. If the `rep` runs lead to different designs, then the tuning parameters of the algorithm can be changed. One can also use use `rep="change"` which runs the algorithm 27 times with different starting and `maxstep` values. An additional example illustrating an instance where Kozak's algorithm does not reach a global minimum is presented in the Appendix.

With $N_u$ unique $X$-values, there are approximately $\binom{N_u-1}{L-1}$ possible sets of stratum boundaries. If this number is smaller than `minsol` all the possible sets of strata are tried, rather than carrying out a random search. The default value is `minsol=1000`. The elements `maxstep`, `maxstill`, `minsol` and `rep` belong to the `algo.control` argument. In Table 7, we were unable to improve the geometric stratified design for the UScities population. The command to run

Kozak's algorithm 27 times with various tuning parameters is given below.

```
> data(UScities)
> pop2LHrep <- strata.LH(x = UScities, CV = 0.0145, Ls = 5,
  alloc = c(0.5, 0, 0.5), takeall = 0, algo = "Kozak",
  algo.control = list(rep = "change"))
```

This command takes a few seconds to run and yields a stratifed design with $n = 100$, similar to that presented in Table 1 for the UScities.

### 3.4 Designs with a predetermined sample size *n*

With Kozak's algorithm it is possible to find the boundaries that minimize the CV of $\bar{y}_s$ for a fixed sample size $n$ rather than minimizing $n$ for a predetermined CV. As an example we revisit the stratified designs of Table 1. The geometric boundaries are used as initial values and the default Kozak algorithm is run. The R-code for the Debtors population is given below.

```
> pop1k <- strata.LH(x = Debtors, initbh = pop1$bh, n = 100,
  Ls = 5, alloc = c(0.5, 0, 0.5), algo = "Kozak")
```

The CVs of the estimator of $\bar{y}_s$ obtained with the optimal stratified designs are 3.12%, 1.43%, 1.72%, and 1.04% for the four populations as compared with 3.59%, 1.45%, 1.83%, and 1.07% in Table 1. Thus the iterative algorithm allowed to reduce the CVs.

## 4. Stratification with anticipated moments

A difference between the stratification variable $X$ and the survey variable $Y$ can be accounted for by having a model for the conditional distribution of $Y$ given $X$. In *stratification*, there is a log-linear model where

$$Y = \exp(\alpha)X^{\beta}\exp(\sigma\epsilon),$$

and an heteroscedastic linear model with

$$Y = \alpha + \beta X + \sigma\epsilon X^{\gamma}, \qquad (2)$$

and $\alpha$, $\beta$, and $\gamma$ are real parameters specified by the user and $\epsilon$ is a $N(0, 1)$ random variable. A random replacement model (Rivest 1999) is also available and stratum specific mortality rates (Baillargeon, Rivest and Ferland 2007) can be added to the log-linear model.

Under these models, the anticipated mean of $Y$ for the units classified in stratum $h$, with $X \in [b_{h-1}, b_h)$ are

$$\bar{Y}_h = \frac{1}{N_h}\sum_{b_{h-1}\le X_i<b_h} E(Y \mid X_i)$$

while the anticipated variance is

$$S^2_{yh} = \frac{1}{N_h}\sum_{b_{h-1}\le X_i<b_h}\{E(Y \mid X_i) - \bar{E}(Y \mid X)_h\}^2 + \frac{1}{N_h}\sum_{b_{h-1}\le X_i<b_h}\mathrm{Var}(Y \mid X_i)$$

where $\bar{E}(Y \mid X)_h$ is the average of the predicted values of $Y$ for the units in stratum $h$. In `strat.cumrootf`, `strata.geo` and `strata.bh` these expressions are used to evaluate the sampling properties of $\bar{y}_s$ while in `strata.LH`, the minimization of (1) is carried out with anticipated moments. In `strata.LH` the stratum boundaries depend on the model for the relationship between $X$ and $Y$; they do not for the other `strata` functions.

### 4.1 An example with the *MU*284 Swedish municipalities

In Section 2.5 two stratified sampling plans were derived for the *MU*284 population with *REV*84 as stratification variable. The R-code that follows investigates the performance of these sampling designs for the variable *RMT*85. The vector `ord` contains the position of the order statistics of the *REV*84 variable; thus `Y[ord]` is the vector of the *RMT*85 variable, ordered by increasing *REV*84 -value.

```
> data(Sweden)
> X <- Sweden$REV84
> Y <- Sweden$RMT85
> ord <- order(X)
> geo_rmt <- var.strata(geo_cer, y = Y[ord])
> cum_rmt <- var.strata(cum_cer, y = Y[ord])
> c(geo_rmt$RRMSE,cum_rmt$RRMSE)

[1] 0.06889558 0.07368794
```

In section 2.4, the CVs of the estimator $\bar{y}_s$ for the stratification variable *REV*84 were less than 5% for the cum$\sqrt{f}$ and the geometric designs. When estimating the mean of *RMT*85, the CVs are larger than 6%. This emphasizes that calculating sample sizes with a stratification variable underestimate the $n$ needed to reach the target CV for a different survey variable. These results are reported in the first two designs of Table 8. Table 8 also shows the optimal design calculated by applying Kozak's algorithm to the *REV*84 variable, assuming $Y = X$.

Following Rivest (2002), a log-linear model is fitted for the relationship between the two variables. As shown in Figure 2, there are outliers and the following R-code estimates the parameters of the log-linear model by discarding the municipalities with extreme $X / Y$ quantiles. The 18 discarded municipalities are represented by a star in Figure 2. The R-code for fitting the model to the non outliers follows.

```
> keep <- (X/Y > quantile(X/Y, 0.03))&(X/Y < quantile(X/Y, 0.97))
> reg <- lm(log(Y)[keep] ~ log(X)[keep])
> coef(reg)

(Intercept) log(X)[keep]
-3.153025   1.058355

> summary(reg)$sigma

[1] 0.25677
```

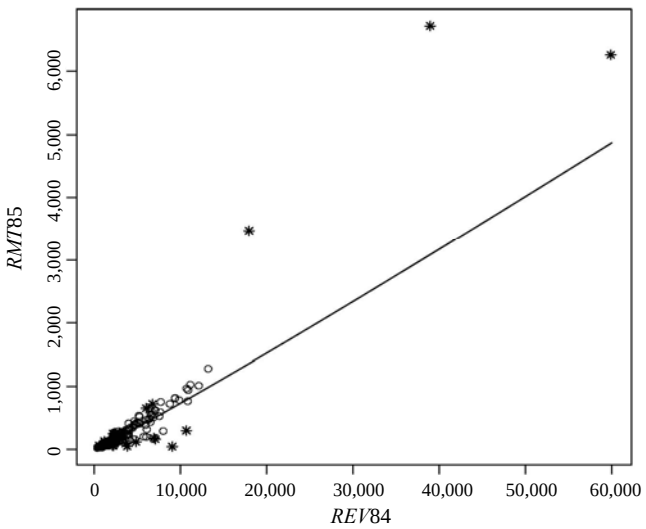


**Figure 2 Plot of *RMT*85 by *REV*84 from the data set Sweden**

The following code stratifies the *MU*284 population on *REV*84 using the cum $\sqrt{f}$ and the geometric method. The allocation is however carried out with anticipated moments calculated with the log-linear regression model of *RMT*85 on *REV*84. The strata of these two designs are the same as those calculated earlier. The model affects only the anticipated CV. It is not so for the optimal design where the anticipated moments are used in the stratification algorithm. Kozak's algorithm might fail to find the global minimum $n$ value when using anticipated moments; thus we use the bounds calculated with $Y = X$ as starting values.

```
> geo_cer.m <- strata.geo(x = X[ord], CV = 0.05, Ls = 4,
  alloc = c(0.35, 0.35, 0), model = "loglinear",
  certain = (length(X) - 2):length(X), model.control =
  list(beta = 1.058355, sig2 = 0.25677^2))
> geo_cer.var <- var.strata(geo_cer.m, y = Y[ord])
> cum_cer.m <- strata.cumrootf(x = X[ord], nclass = 50,
  CV = 0.05, Ls = 4, alloc = c(0.35, 0.35, 0),
  certain = (length(X) - 2):length(X), model = "loglinear",
  model.control = list(beta = 1.058355, sig2 = 0.25677^2))
> cum_cer.var <- var.strata(cum_cer.m, y = Y[ord])
> LH <- strata.LH(x = X, CV = 0.05, Ls = 5,
  alloc = c(0.35, 0.35, 0), takeall = 1)
> LH.var <- var.strata(LH, y = Y)
> LH_m <- strata.LH(x = X, CV = 0.05, Ls = 5,
  initbh = LH$bh, alloc = c(0.35, 0.35, 0), takeall = 1,
  model = "loglinear", model.control = list(beta = 1.058355,
  sig2 = 0.25677^2))
> LH_m.var <- var.strata(LH_m, y = Y)
```

In Table 8, sample sizes calculated with anticipated moments give CVs smaller than 5% for estimating the mean *RMT*85 variable. The optimal LH design requires a $n$ slightly smaller than the other two. Accounting for $Y \neq X$ when minimizing (1) gives a larger take-all stratum since its size increased from 4 to 5 when using the anticipated moments.

Finally observe that the arguments `model` and `model.control` can be used with `var.strata`. For the geometric design considered in this section, one can get results very similar to those obtained with the argument `y=Y`. As shown below, the model yields a CV of 6.894% as compared with 6.890% obtained with the original *RMT*85-variable. For the cum $\sqrt{f}$ method the model CV is 7.282% as compared to 7.369% found earlier while for the Lavallée Hidiroglou algorithm these two values are 7.080% and 7.110%.

```
> geo_rmt2 <- var.strata(geo_cer, model = "loglinear",
  model.control = list(beta = 1.058355, sig2 = 0.25677^2))
> geo_rmt2$RRMSE

[1] 0.0689368
```

**Table 8**
**Three stratified designs for estimating the mean *RMT*85 with *REV*84 as the stratification variable**

| Model | Method | | 1 | 2 | 3 | 4 | 5 | *n* anticip. | CV |
|---|---|---|---|---|---|---|---|---|---|
| $Y = X$ | cum $\sqrt{f}$ | $N_h$ | 127 | 79 | 46 | 29 | 3 | | |
| | | $n_h$ | 3 | 4 | 4 | 5 | 3 | 19 | 7.37 |
| | geometric | $N_h$ | 42 | 116 | 88 | 35 | 3 | | |
| | | $n_h$ | 2 | 5 | 7 | 7 | 3 | 24 | 6.89 |
| | LH | $N_h$ | 120 | 82 | 45 | 33 | 4 | | |
| | | $n_h$ | 3 | 4 | 4 | 5 | 4 | 20 | 7.11 |
| loglinear | cum $\sqrt{f}$ | $N_h$ | 127 | 79 | 46 | 29 | 3 | | |
| | | $n_h$ | 6 | 8 | 9 | 10 | 3 | 36 | 4.78 |
| | geometric | $N_h$ | 42 | 116 | 88 | 35 | 3 | | |
| | | $n_h$ | 3 | 8 | 13 | 13 | 3 | 40 | 4.74 |
| | LH | $N_h$ | 121 | 81 | 45 | 32 | 5 | | |
| | | $n_h$ | 6 | 7 | 7 | 9 | 5 | 34 | 4.90 |

### 4.2 Anderson, Kish and Cornell (1976) example with the bivariate normal distribution

Anderson *et al.* (1976) investigated the optimal stratification for $Y$ based on $X$ when $(X, Y)$ has a bivariate normal distribution with correlation $\rho$. Thus model (2) holds with $\alpha = \gamma = 0$, $\beta = \rho$, and $\sigma^2 = 1 - \rho^2$ where $X$ has a $N(0, 1)$ distribution. To reproduce Anderson *et al.* (1976) results, we generate a population of size $N = 10^5$ from a $N(0, 1)$ distribution and select `model="linear"` (as in Section 3.1 a mean of 10 was used to prevent $X$ from being negative). For a linear model, only Kozak's algorithm works. Given the special nature of the problem, the `maxstep` parameter is set to 20 and only one repetition (`rep=1`) of the algorithm is run. When there is no take-all stratum, the optimal stratum boundaries are independent of the CV, as in Section 3.1. We used CV = 0.01 in the calculations.

```
> x <- rnorm(1e+05, 10)
> bi3a <- strata.LH(x = x, CV = 0.01, Ls = 3, takenone = 0,
  model = "linear",
  model.control = list(beta = 0.25, sig2 = 1 - 0.25^2,
  gamma = 0), algo.control = list(maxstep = 20, rep = 1))
> bi3a$bh - 10

[1] -0.619354 0.604198
```

In Table 9, *stratification*'s results are equal to Anderson's *et al.* (1976) findings up to nearly two decimals. This highlights the flexible nature of the package; it can find the optimal stratified design for any distribution of the stratification variable and for some general models for the conditional distribution of $Y$ given $X$.

**Table 9**
**Comparison of Anderson *et al.* (1976) optimal stratum boundaries with the approximate boundaries obtained with *stratification***

| | | ***stratification*'s results** | | | | **Anderson *et al.*'s results** | | | |
|---|---|---|---|---|---|---|---|---|---|
| ***L*** | **\|ρ\|** | **1** | **2** | **3** | **4** | **1** | **2** | **3** | **4** |
| 3 | 0.250 | -0.619 | 0.604 | | | -0.61 | 0.61 | | |
| | 0.950 | -0.591 | 0.568 | | | -0.58 | 0.58 | | |
| | 0.990 | -0.571 | 0.549 | | | -0.56 | 0.56 | | |
| 4 | 0.250 | -0.984 | 0.004 | 0.985 | | -0.98 | 0.00 | 0.98 | |
| | 0.950 | -0.930 | 0.009 | 0.942 | | -0.93 | 0.00 | 0.93 | |
| | 0.990 | -0.902 | -0.001 | 0.895 | | -0.90 | 0.00 | 0.90 | |
| 5 | 0.250 | -1.245 | -0.377 | 0.387 | 1.251 | -1.24 | -0.38 | 0.38 | 1.24 |
| | 0.950 | -1.187 | -0.358 | 0.372 | 1.197 | -1.19 | -0.37 | 0.37 | 1.19 |
| | 0.990 | -1.136 | -0.344 | 0.353 | 1.144 | -1.14 | -0.35 | 0.35 | 1.14 |

## 5. Additional features

Baillargeon and Rivest (2009) considered additional aspects of a stratified design, namely stratum specific anticipated non-response rates and the addition of a take-none stratum with a null sample size. This section discusses briefly how these additional items are handled in *stratification*. Non-response needs to be accounted for when optimizing for $n$. A take-none stratum makes $\overline{y}_s$ biased; in this case the precision target is specified in terms of a Relative Root Mean Squared Error (RRMSE) rather than a CV. Formula (4.3) of Baillargeon and Rivest (2009) provides a generalization of (1) that includes these two features. This is the formula used for calculating sample sizes in the optimization procedure.

### 5.1 Non-response

Non-response can be corrected *a posteriori*, by dividing the no non-response stratum sample sizes by the response rates. This is illustrated in the following R-code that considers the MRTS variable, representative of Statistics Canada Monthly Retail Trade Survey. *Post hoc* non-response corrections are implemented in the `var.strata` function with the argument `rh.postcorr=TRUE`. An alternative is to consider response rates when allocating the sample to the strata. They can be specified in a `strata` function with the argument `rh=`. This approach penalizes strata with a high non-response; it typically yields a smaller $n$ value than the *a posteriori* corrections. This is illustrated in the cum $\sqrt{f}$ portion of Table 10. With four strata and response rates of 0.8, 0.8, 0.9, 1, the *a posteriori* correction needs $n = 445$ to reach the target CV for the MRTS variable, as compared with $n = 444$ for an allocation that takes non-response into account.

```
> data(MRTS)
> cum <- strata.cumrootf(x = MRTS, nclass = 500, CV = 0.01,
  Ls = 4, alloc = c(0.5, 0, 0.5))
> cum.var <- var.strata(cum, rh = c(0.8, 0.8, 0.9, 1))
> cum.post <- var.strata(cum, rh = c(0.8, 0.8, 0.9, 1),
  rh.postcorr = TRUE)
> cum_rh <- strata.cumrootf(x = MRTS, nclass = 500, CV = 0.01,
  Ls = 4, alloc = c(0.5, 0, 0.5), rh = c(0.8, 0.8, 0.9, 1))
```

Non-response can also be accounted for when constructing an optimal sampling design, either *a posteriori* or in the stratum construction. These two approaches are implemented for the MRTS population in the following R-code. The higher non-response rates for the small units penalize the first stratum which is smaller when non-response is accounted for in the stratification algorithm, as can be seen in Table 10. Still accounting for non-response in the stratum construction gives a smaller $n$-value than an *a posteriori* correction. Table 3 of Baillargeon and Rivest (2009) presents additional examples, including both anticipated moments and non-response, of the construction of stratified designs for the MRTS population.

```
> LH <- strata.LH(x = MRTS, CV = 0.01, Ls = 4,
  alloc = c(0.5, 0, 0.5), takeall = 1)
> LH.var <- var.strata(LH, rh = c(0.8, 0.8, 0.9, 1))
> LH.post <- var.strata(LH, rh = c(0.8, 0.8, 0.9, 1),
  rh.postcorr = TRUE)
> LH_rh <- strata.LH(x = MRTS, CV = 0.01, Ls = 4,
  alloc = c(0.5, 0, 0.5), takeall = 1, rh = c(0.8, 0.8, 0.9, 1))
```

**Table 10**
**Two examples of non-response correction: Either *a posteriori* (post) or when constructing the design**

| **Method** | **rh** | | **1** | **2** | **3** | **4** | ***n*** | **anticip. CV** |
|---|---|---|---|---|---|---|---|---|
| cum $\sqrt{f}$ | none | $N_h$ | 778 | 742 | 355 | 125 | | |
| | | $n_h$ | 87 | 90 | 88 | 125 | 390 | 1.11 |
| | | $n_h^{post}$ | 109 | 113 | 98 | 125 | 445 | 1.00 |
| | given | $N_h$ | 778 | 742 | 355 | 125 | | |
| | | $n_h$ | 105 | 108 | 106 | 125 | 444 | 1.00 |
| LH | none | $N_h$ | 774 | 675 | 374 | 177 | | |
| | | $n_h$ | 77 | 65 | 60 | 177 | 379 | 1.11 |
| | | $n_h^{post}$ | 96 | 81 | 67 | 177 | 421 | 1.00 |
| | given | $N_h$ | 675 | 677 | 449 | 199 | | |
| | | $n_h$ | 70 | 69 | 80 | 199 | 418 | 1.00 |

### 5.2 Take-none stratum

A take-none stratum with a null sample size might be advantageous when the population has small units with $Y$-values close to 0. The precision of $\overline{y}_s$ is then measured by the mean squared error, $\mathrm{Var}(\overline{y}_s) + (T_{0y}/N)^2$, where $T_{0y}$ is

the anticipated $Y$-total in the take-none stratum. Setting `takenone=1` in the `strata.LH` function constructs an optimal design with a take-none stratum. Baillargeon and Rivest (2009) showed that Sethi's algorithm does not work in this case and that Kozak's algorithm should be used. When a take-none stratum is used, a rough bias correction can be implemented by dividing $\overline{y}_s$ by the proportion of the total of the $X$ variable in the take some strata. Thus the bias penalty in the mean square error might be too stringent and an alternative measure of precision, such as $\text{Var}(\overline{y}_s) + (p \times T_{0y}/N)^2$, could be used in the stratification algorithm where $p$ is a number in $(0, 1)$. This smaller bias penalty can be implemented by setting the argument `bias.penalty` equal to $p$. The following R-code constructs three optimal stratified designs for the MRTS population, with and without a take-none stratum; the default full bias penalty is compared to a reduced penalty with $p = 0.5$.

```
> data(MRTS)
> notn <- strata.LH(x = MRTS, CV = 0.1, Ls = 3,
  alloc = c(0.5, 0, 0.5))
> tn1 <- strata.LH(x = MRTS, CV = 0.1, Ls = 3,
  alloc = c(0.5, 0, 0.5), takenone = 1)
> tn0.5 <- strata.LH(x = MRTS, CV = 0.1, Ls = 3,
  alloc = c(0.5, 0, 0.5), takenone = 1, bias.penalty = 0.5)
```

The sample sizes $n$ for the three designs are given in Table 11. Including a take-none stratum with a full bias penalty reduces $n$, from 22 to 16; for this design the take-none stratum accounts for 3% of the total of the X-variable. Reducing the biais penalty to $p = 0.5$ increases the size of the take-none stratum and reduces $n$. Additional illustrations are given in Table 2 of Baillargeon and Rivest (2009). They show that the size of a take-none stratum typically decreases with the target RRMSE. For the MRTS example, the addition of a take-none stratum diminishes the $n$-value substantially while for others it does not change the design.

**Table 11**
**Sample sizes for three optimal stratified designs for the MRTS population**

| | | | |
|---|---|---|---|
| **takenone** | 0 | 1 | 1 |
| **bias.penalty** | NA | 1 | 0.5 |
| ***n*** | 22 | 16 | 13 |
| **% $T_x$** | 0 | 3 | 9 |

## 6. Conclusion

The R-package *stratification* offers flexible methods for the construction of a stratified sampling design using a univariate stratification variable such as a measure of size in a business survey. Several methods are available to determine the stratum boundaries and the stratum sample sizes. *stratification* allows the investigation of features such as a take-all stratum, a take-none stratum, the extent of the discrepancy between X and Y, and a stratum specific non-response.

## Acknowledgements

We are grateful to S. Er, E. Gagnon, M. Kozak, and J. Stardom for constructive comments on the package and to the Canada Research Chair on Statistical Sampling and Data Analysis and the Natural Sciences and Engineering Research Council of Canada for their financial support. This research was supported by U.S. National Science Foundation grant SES-0751671.

## 7. Appendix

### 7.1 More details on Kozak's algorithm

As described in Section 3.3 Kozak's algorithm uses a random search. Besides decreasing the optimization criterion, either the $n$-value or the *RRMSE* of $\overline{y}_s$, *stratification* requires that the take-some strata contain at least `minNh` units and that they have positive sample sizes, for the new boundary to be admissible. The default is `minNh=2`. A non random, Kozak's algorithm is also available with `method="modified"` in the `algo.control` argument. It tries all the possible changes at one iteration and picks the one that gives the largest drop of the optimization criterion. It is slower than Kozak's algorithm without improving the detection of the global minimum of the optimization criterion. Therefore, it will not be discussed any further.

To illustrate the complete enumeration of all possible solutions mentioned in Section 3.3, consider the `USbanks` data set. It contains 357 values, but only 200 unique values. If one wishes to stratify this population in two strata, only $\binom{200-1}{2-1} = 199$ solutions are possible. The following command performs a complete enumeration of the possible solutions:

```
> enum <- strata.LH(x = USbanks, CV = 0.05, Ls = 2,
  alloc = c(0.5, 0, 0.5))
```

These solutions, with their associated optimization criteria value, can be found in `enum$sol.detail`. Only the solutions fulfilling the admissibility constraints mentioned above are included in `enum$sol.detail`.

When running Kozak's algorithm, the initial boundary values might fail to meet the admissibility constraints; the algorithm might not be able to move at all. In such a case, the initial boundaries are replaced by robust ones. The robust boundaries give an empty take-none stratum if such a stratum is requested, take-all strata as small as possible, and take-some strata with approximately the same number of unique $X$-values.

Consider once again the example of Section 3.2 with the `UScities` data set, where Kozak's algorithm reached a local minimum with the default arguments. With geometric initial boundaries, Kozak's algorithm converges rapidly to what appears to be a global minimum.

```
> LH_init <- strata.LH(x = UScities, initbh = pop2$bh,
  n = 100, Ls = 5, alloc = c(0.5, 0, 0.5), takeall = 0,
  algo.control = list(rep = 1))
> LH_init$iter.detail

      b1    b2    b3   b4         opti   step   iter   run
1   18.5  33.5  59.5  107   0.01444981      0      0     1
2   20.5  33.5  59.5  107   0.01435576      2      2     1
3   19.5  33.5  59.5  107   0.01434272     -1     10     1
4   19.5  33.5  58.0  107   0.01432714     -1     12     1
5   19.5  31.5  58.0  107   0.01431013     -2     13     1
6   19.5  32.5  58.0  107   0.01430163      1     63     1

> LH_init$niter
[1] 163
```

The output element `LH_init$iter.detail` contains information about the initial boundaries and the 5 iterations with a change of boundaries only. A total of 163 iterations were needed for the algorithm to converge. The geometric initial boundaries are very close to the optimal solutions. A local minimum can also be avoided by changing some of the algorithm's parameters. The following R-code allows larger steps (`maxstep=20`) and increases the maximal number of iterations (`maxstill=1000`) and the number of repetitions of the algorithm (`rep=20`).

```
> LH_param <- strata.LH(x = UScities, n = 100, Ls = 5,
  alloc = c(0.5, 0, 0.5), takeall = 0, algo.control =
  list(maxstep = 20, maxstill = 1000, rep = 20))
```

The results for the 20 repetitions are reported in `LH_param$rep.detail` and summarized in Table 12. The solution obtained with the geometric initial boundaries is reached 9 times out of 20.

**Table 12**
**Solutions found by Kozak's algorithm for 20 repetitions**

| CV | B1 | B2 | B3 | B4 | frequency |
|---|---|---|---|---|---|
| 0.0143 | 19.50 | 32.50 | 58.00 | 107.00 | 9 |
| 0.0167 | 16.50 | 23.50 | 37.50 | 78.00 | 5 |
| 0.0167 | 15.50 | 22.50 | 35.50 | 73.00 | 6 |

Figure 3 shows how larger steps help the algorithm to reach the global minimum (CV = 0.0143), compared to a run of the algorithm with the default arguments (dotted lines, CV = 0.0167).

### 7.2 R package *stratification* summary table

This appendix provides a quick reference for the `R` package *stratification*. Table 13 lists the five functions in *stratification* and their arguments. The following notes complete the table.

(1) According to the general allocation scheme (Hidiroglou and Srinath 1993). The stratum sample sizes are proportional to $N_h^{2q_1}\overline{Y}_h^{2q_2}S_{yh}^{2q_3}$.

(2) The default value of `initbh` is the set of arithmetic starting points of Gunning and Horgan (2007), see Section 3.3. If `takenone=1` and `initbh` is of size `Ls-1`, the initial boundary of the take-none stratum is set to the first percentile of `X`. If this first percentile is equal to the minimum value of `X`, this initial boundary would lead to an empty take-none stratum. In that case, the initial boundary of the take-none stratum is rather set to the second smallest value of `X`.

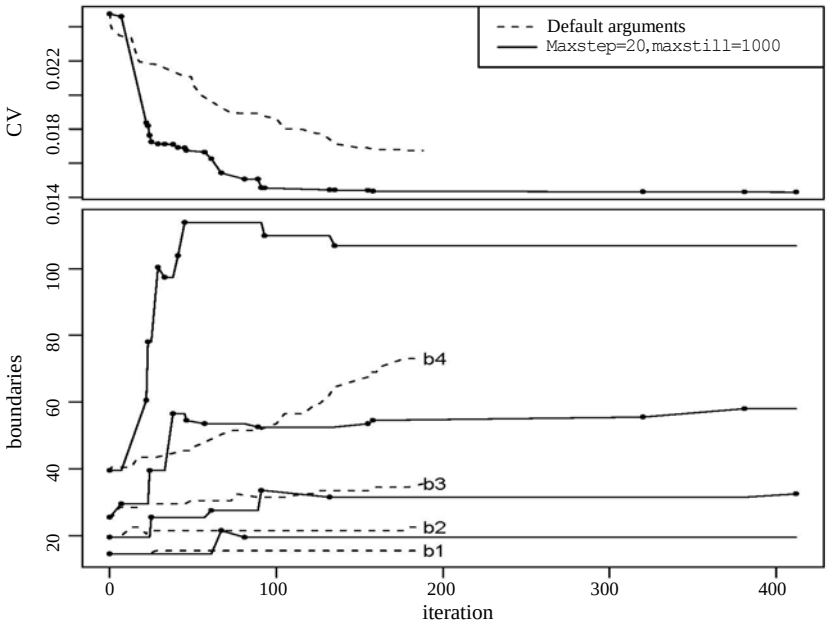


**Figure 3 Iterations histories for two runs of Kozak's algorithm**

(3) The elements to specify in the `algo.control` argument depend on the algorithm. The following table shows the elements used by each algorithm and their default values. See `help(strata.LH)` for a complete description of every element.

| Algorithm | maxiter | method | minNh | maxstep | maxstill | rep | minsol |
|---|---|---|---|---|---|---|---|
| Sethi | 500 | - | - | - | - | - | - |
| Original Kozak | 10,000 | "original" | 2 | 3 | 100 | 3 | 1,000 |
| Modified Kozak | 3,000 | "modified" | 2 | 3 | - | - | 1,000 |

(4) The elements of the `model.control` argument depend on the model:

- `loglinear` model with mortality:

$$Y = \begin{cases} \exp(\alpha + \texttt{beta}\,\log(X) + \texttt{epsilon}) & \\ \qquad\qquad \text{with probability } p_h & \\ 0 \qquad\qquad \text{with probability } 1-p_h & \end{cases}$$

where $\texttt{epsilon} \sim N(0, \texttt{sig2})$ is independent of $X$. The parameter $p_h$ is specified through `ph`, `ptakenone` and `pcertain`.

- heteroscedastic `linear` model :

$$Y = \texttt{beta}\,X + \texttt{epsilon}$$

where

$$\texttt{epsilon} \sim N(0, \texttt{sig2}\,X^{\texttt{gamma}}).$$

- `random` replacement model:

$$Y = \begin{cases} X & \text{with probability } 1 - \texttt{epsilon} \\ X\text{new} & \text{with probability } \texttt{epsilon} \end{cases}$$

where $X$new is a random variable independent of $X$ with the same distribution as $X$.

The following table presents `model.control` default values according to the model.

| model | beta | sig2 | ph | ptakenone | pcertain | gamma | epsilon |
|---|---|---|---|---|---|---|---|
| "loglinear" | 1 | 0 | rep(1,Ls) | 1 | 1 | - | - |
| "linear" | 1 | 0 | - | - | - | 0 | - |
| "random" | - | - | - | - | - | - | 0 |

**Table 13**
**R package *stratification* summary table**

| argument | Strata.cumrootf | Strata.geo | Strata.LH | Strata.bh | Var.strata | description | format | default |
|---|---|---|---|---|---|---|---|---|
| x | • | • | • | • | | stratification variable | vector | none (x is mandatory) |
| n | • | • | • | • | | target total sample size | scalar | none (n or CV is mandatory) |
| CV | • | • | • | • | | target CV or RRMSE | scalar | none (n or CV is mandatory) |
| Ls | • | • | • | • | | number of sampled strata | scalar | 3 |
| alloc | • | • | • | • | | allocation specification (1) | list (q1,q2,q3) where qi≥ 0 | Neyman (q1=q3=0.5, q2=0) |
| certain | • | • | • | • | | x–indices for units sampled with certainty | vector | NULL (no certainty stratum) |
| nclass | • | | | | | number of bins | scalar | *min*(10*L*, *N*) |
| bh | | | | • | | strata boundaries | vector | none (bh is mandatory) |
| takeall.adjust | | | | • | | indicator of adjustment for take-all strata | True or False | FALSE (no adjustment) |
| takeall | | | • | • | | number of take-all strata | one of {0, 1,..., Ls − 1} | 0 |
| initbh | | | • | | | initial strata boundaries (2) | vector | equidistant boundaries |
| algo | | | • | | | algorithm identification | "Kozak" or "Sethi" | "Kozak" |
| algo.control | | | • | | | algorithm's parameters specification (3) | list (maxiter, method, minNh, maxstep, maxstill, rep) | depends on algo |
| strata | | | | | • | stratification scheme | strata object | none (strata is mandatory) |
| y | | | | | • | study variable | vector | NULL (model given instead) |
| model | • | • | • | • | • | model identification | "none", "loglinear", "linear"* or "random"* → | "none" (*unavailable with Sethi's algo) |
| model.control | • | • | • | • | • | model's parameter specification (4) | list (beta, sig2, ph, ptakenone, gamma, epsilon) | depends on model, but equivalent to model="none" |
| rh | • | • | • | • | • | anticipated response rates | scalar or vector | rep(1,Ls) or rh from strata |
| rh.postcorr | | | | | • | indicator of posterior correction for non-response | TRUE or FALSE | FALSE (no correction) |
| takenone | | | • | • | | number of take-none strata | 0 or 1 | 0 |
| bias.penalty | | | • | • | | penalty for the bias | scalar | 1 |

## References


Anderson, D.W., Kish, L. and Cornell, R.G. (1976). Quantifying gains from stratification for optimum and approximately optimum strata using a bivariate normal model. *Journal of the American Statistical Association*, 71, 887-892.

Baillargeon, S., Rivest, L.-P. and Ferland, M. (2007). Stratification en enquêtes entreprises : une revue et quelques avancées. *Proceedings of the Survey Methods Section, Statistical Society of Canada* (www.ssc.ca/survey/documents/SSC2007_S_Baillargeon.pdf).

Baillargeon, S., and Rivest, L.-P. (2009). A general algorithm for univariate stratification. *International Statistical Review*, 77, 331-344.

Cochran, W.G. (1961). Comparison of methods for determining stratum boundaries. *Bulletin of the International Statistical Institute*, 32, 345-358.

Cochran, W.G. (1977). *Sampling Techniques. Third Edition*. New York: John Wiley & Sons, Inc.

Dalenius, T., and Hodges, J.L., Jr. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, 54, 88-101.

Dayal, S. (1985). Allocation of sample using values of auxiliary characteristics. *Journal of Statistical Planning and Inference*, 11, 321-328.

Detlefsen, R.E., and Veum, C.S. (1991). Design issues for the Retail Trade Sample Surveys of the U.S. Bureau of the Census. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 214-219.

Gunning, P., and Horgan, J.M. (2004). A new algorithm for the construction of stratum boundaries in skewed populations. *Survey Methodology*, 30, 159-166.

Gunning, P., and Horgan, J.M. (2007). Improving the Lavallée and Hidiroglou algorithm for stratification of skewed populations. *Journal of Statistical Computation and Simulation*, 77, 277-291.

Hidiroglou, M.A. (1986). The construction of a self-representing stratum of large units in survey design. *The American Statistician*, 40, 27-31.

Hidiroglou, M.A., and Srinath, K.P. (1993). Problems associated with designing subannual business surveys. *Journal of Business and Economic Statistics*, 11, 397-405.

Khan, M.G.M., Nand, N. and Ahmad, N. (2008). Determining the optimum strata boundary points using dynamic programming. *Survey Methodology*, 34, 205-214.

Kozak, M. (2004). Optimal stratification using random search method in agricultural surveys. *Statistics in Transition*, 6, 797-806.

Kozak, M., and Verma, M.R. (2006). Geometric versus optimization approach to stratification: A comparison of efficiency. *Survey Methodology*, 32, 157-163.

Lavallée, P., and Hidiroglou, M. (1988). On the stratification of skewed populations. *Survey Methodology*, 14, 33-43.

McEvoy, R.H. (1956). Variation in bank asset portfolios. *The Journal of Finance*, 11(4), 463-473.

Rivest, L.-P. (1999). Stratum jumpers: Can we avoid them?. *ASA Proceedings of the Section on Survey Research Methods*, American Statistical Association, (Alexandria, VA), 64-72.

Rivest, L.-P. (2002). A generalization of the Lavallée and Hidiroglou algorithm for stratification in business surveys. *Survey Methodology*, 28, 191-198.

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

Sethi, V.K. (1963). A note on the optimum stratification of populations for estimating the population means. *Australian Journal of Statistics*, 5, 20-33.

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

Slanta, J., and Krenzke, T. (1996). Applying the Lavallée and Hidiroglou method to obtain stratification boundaries for the Census Bureau's Annual Capital Expenditure Survey. *Survey Methodology*, 22, 65-75.

Sweet, E.M., and Sigman R.S. (1995). Evaluation of model-assisted procedures for stratifying skewed populations using auxiliary data. *U.S. Bureau of the Census* (www.census.gov/srd/papers/pdf/sm95-22.pdf).


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

Survey Methodology, June 2011
Vol. 37, No. 1, pp. 67-74
Statistics Canada, Catalogue No. 12-001-X



# Replication variance estimation under two-phase sampling

**Jae Kwang Kim and Cindy Long Yu** [1]

Abstract

In two-phase sampling for stratification, the second-phase sample is selected by a stratified sample based on the information observed in the first-phase sample. We develop a replication-based bias adjusted variance estimator that extends the method of Kim, Navarro and Fuller (2006). The proposed method is also applicable when the first-phase sampling rate is not negligible and when second-phase sample selection is unequal probability Poisson sampling within each stratum. The proposed method can be extended to variance estimation for two-phase regression estimators. Results from a limited simulation study are presented.

Key Words: Double sampling; Jackknife; Regression estimator; Reweighted expansion estimator.

## 1. Introduction

Two-phase sampling, first introduced by Neyman (1938) and sometimes called double sampling, is a cost effective technique in survey sampling. It is typically used when it is very expensive to collect data on the variables of interest, but it is relatively inexpensive to collect data on variables that are correlated with the variables of interest. Two-phase sampling has application in different forms (*e.g.*, Rao 1973; Cochran 1977; Breidt and Fuller 1993; Rao and Sitter 1995; Hidiroglou and Särndal 1998; Fuller 1998; Hidiroglou 2001; Fuller 2003). Two-phase sampling for stratification refers to the situation where the observation from the first-phase sample is used to make a stratification for the second-phase sampling. By selecting the first-phase sample for stratification purpose, two-phase sampling is a useful tool when there is no sampling frame available for stratification at the beginning. For example, in forest surveys, it is very difficult and expensive to travel to remote areas to make on-ground determinations. However, aerial photographs are relatively inexpensive, and determinations on, say, forest type from aerial photos are strongly correlated with ground determinations and can be used to stratify the first phase sample.

Replication variance estimation is very popular in complex surveys. Rust and Rao (1996) and Wolter (2007) provide comprehensive overviews on this topic. The replication method does not require the computation of the partial derivative of the Taylor expansion and the user can easily produce variance estimates without knowing the sampling design that was used to collect the data. Furthermore, this tendency is increasing because of confidentiality issues (Lu and Sitter 2006). Once the replication weights are provided, the design information such as stratum identifier is not needed for the user's analysis.

There are two commonly used estimators of the population mean under two phase sampling: the double expansion estimator (DEE) and the reweighted expansion estimator (REE), named by Kott and Stukel (1997). In general the REE is more efficient than the DEE in the situation of two-phase sampling for stratification when the $y$'s within a stratum are homogeneous. Variance estimation for two-phase sampling is a challenging practical problem, and replication variance estimation is of interest among practitioners. Rao and Shao (1992) proposed a consistent jackknife variance estimator for the REE in the context of hot deck imputation treating the respondents as the second-phase sample. Kott and Stukel (1997) considered the same problem and concluded that the jackknife variance estimator works well for the REE if the first-phase sampling rate is negligible. The sampling rate, or the sampling fraction, $f_1 = nN^{-1}$ is called negligible if $f_1$ converges to zero under the asymptotic setup described in Section 2. Binder, Babyak, Brodeur, Hidiroglou and Jocelyn (2000) studied variance estimation for a similar two-phase sample design using the Taylor linearization method. Kim *et al.* (2006, KNF) provided a rigorous investigation of the replication method and considered replication for other types of estimators. The KNF method has been developed mainly under the situation where the first-phase sampling rate is negligible and the second-phase sampling is a stratified random sampling. If the first-phase sampling rate is not negligible, additional replicates are needed to get consistent variance estimates.

In this paper, we propose a new replication method for variance estimation under two-phase sampling. The proposed method is an extension of the KNF method to cover the situation where the first-phase sampling rate is not necessarily negligible. Unlike the KNF method, the proposed method does not require additional replicates for bias correction in the variance estimation, but does require adjustments in the replication weights. Also, the proposed method is applicable to unequal probability Poisson sampling within

1. Jae Kwang Kim, Department of Statistics, Iowa State University, Ames, Iowa 50011, U.S.A.; Cindy Long Yu, Department of Statistics, Center for Survey Statistics and Methodology, Iowa State University, Ames, Iowa 50011, U.S.A. E-mail: cindyyu@iastate.edu.

second-phase strata, which was not discussed in KNF. Because the proposed method is a replication-based method, it is very easy to implement and can be applied to various types of estimators.

The rest of the paper is organized as follows. In Section 2, the basic setup is introduced, and in Section 3, the proposed method is described. In Section 4, the proposed method is extended to other estimators in two-phase sampling. In Section 5, results from a limited simulation study are presented. Concluding remarks are made in Section 6.

## 2. Basic setup

For better motivation, in this section we simply assume the situation where the first phase is a simple random sample of size $n$ from a finite population of size $N$ and the second phase sampling is a stratified random sample. In section 3, the setup is extended to include any arbitrary measurable sampling in the first phase and unequal probability Poisson sampling within each stratum in the second phase. Using the information obtained from the first-phase sample, it is stratified into $H$ strata for second-phase sampling. In stratum $h$, we have $n_h$ first-phase sample elements and let $A_{h1}$ be the set of indices for the first-phase sample elements in stratum $h$. In the second-phase sampling, a stratified random sample of size $r$ is selected with sample size $r_h (\le n_h)$ in stratum $h$, where $r = \sum_{h=1}^{H} r_h$ and the sampling rate $r_h/n_h$ is fixed for each stratum. To formally discuss the asymptotic theory, we assume a sequence of finite populations, a sequence of first-phase samples, and a sequence of second-phase samples, as described in KNF. In this asymptotic setup, we allow that the second-phase sample size $r$ goes to infinity at the same rate as the first phase sample size $n$, *i.e.*, $r = O(n)$ and $r^{-1} = O(n^{-1})$, and $H$ is fixed. Thus, in the setup of fixed $H$, $r_h^{-1} = O(n^{-1})$.

When the study variable $y_i$ is observed in the second phase sample, the population mean of $y$ is estimated by

$$\bar{y}_{tp} = \frac{1}{n}\sum_{h=1}^{H}\sum_{i\in A_{h2}}\frac{n_h}{r_h}\, y_i,$$

where $A_{h2}$ is the set of indices for the second-phase sample elements that belong to stratum $h$. The variance of $\bar{y}_{tp}$ can be written as

$$\mathrm{Var}(\bar{y}_{tp}) = \left(\frac{1}{n}-\frac{1}{N}\right)S^2 + E\left\{\sum_{h=1}^{H}\left(\frac{n_h}{n}\right)^2\left(\frac{1}{r_h}-\frac{1}{n_h}\right)s_{h1}^2\right\} \quad (1)$$

where $\bar{y}_1 = n^{-1}\sum_{h=1}^{H}\sum_{i\in A_{h1}} y_i$, $S^2 = (N-1)^{-1}\sum_{i=1}^{N}(y_i - \bar{y}_N)^2$, $s_{h1}^2 = (n_h-1)^{-1}\sum_{i\in A_{h1}}(y_i - \bar{y}_{h1})^2$, and $\bar{y}_{h1} = n_h^{-1}\sum_{i\in A_{h1}} y_i$. Using

$$n^{-1}S^2 \doteq E\left\{n^{-1}\sum_{h=1}^{H} w_h[(\bar{y}_{h1}-\bar{y}_1)^2 + s_{h1}^2]\right\}$$

where $w_h = n^{-1}n_h$ and $\doteq$ indicates an approximation ignoring the terms of order $o(n^{-1})$, the variance term (1) is approximated by

$$\mathrm{Var}(\bar{y}_{tp}) \doteq E\left\{n^{-1}(1-f_1)\sum_{h=1}^{H} w_h\,(\bar{y}_{h1}-\bar{y}_1)^2 + \sum_{h=1}^{H}(r_h^{-1}-n_h^{-1}f_1)\,w_h^2\,s_{h1}^2\right\}, \quad (2)$$

where $f_1 = nN^{-1}$.

A consistent estimator of the variance of $\bar{y}_{tp}$ can be derived from (2) by replacing $\bar{y}_{h1}$ and $s_{h1}^2$ by their estimates $\bar{y}_{h2} = r_h^{-1}\sum_{i\in A_{h2}} y_i$ and $s_{h2}^2 = (r_h-1)^{-1}\sum_{i\in A_{h2}}(y_i - \bar{y}_{h2})^2$, respectively. That is, a consistent variance estimator is

$$\hat{V} = n^{-1}(1-f_1)\sum_{h=1}^{H} w_h(\bar{y}_{h2}-\bar{y}_2)^2 + \sum_{h=1}^{H}(r_h^{-1}-n_h^{-1}f_1)\,w_h^2\,s_{h2}^2, \quad (3)$$

where $\bar{y}_2 = \sum_{h=1}^{H} w_h\,\bar{y}_{h2}$. The variance estimator (3) is a linearized variance estimator.

Kott and Stukel (1997) and KNF developed a jackknife variance estimator by successively deleting units from the entire first-phase sample and then adjusting the weights. The full jackknife replicates are

$$\bar{y}_{tp}^{(k)} = \frac{1}{N}\sum_{h=1}^{H}\hat{N}_{h1}^{(k)}\,\bar{y}_{h2}^{(k)} \quad (4)$$

where $k$ is the index of the unit deleted in the jackknife replicate,

$$\frac{1}{N}\hat{N}_{h1}^{(k)} = \sum_{i\in A_{h1}} w_i^{(k)} = \begin{cases}(n-1)^{-1}(n_h-1) & \text{if } k\in A_{h1}\\ (n-1)^{-1}n_h & \text{if } k\notin A_{h1}\end{cases}$$

and

$$\bar{y}_{h2}^{(k)} = \frac{\sum_{i\in A_{h2}} w_i^{(k)} y_i}{\sum_{i\in A_{h2}} w_i^{(k)}} = \begin{cases}(r_h-1)^{-1}(r_h\bar{y}_{h2}-y_k) & \text{if } k\in A_{h2}\\ \bar{y}_{h2} & \text{if } k\notin A_{h2}.\end{cases} \quad (5)$$

The full jackknife variance estimator of the form

$$\hat{V}_J = \sum_{k\in A_1}\frac{n-1}{n}(1-f_1)\,(\bar{y}_{tp}^{(k)}-\bar{y}_{tp})^2, \quad (6)$$

where $\bar{y}_{tp}^{(k)}$ is defined in (4), is asymptotically equivalent to

$$\hat{V}_J \doteq n^{-1}(1-f_1)\sum_{h=1}^{H} w_h(\bar{y}_{h2}-\bar{y}_2)^2 + (1-f_1)\sum_{h=1}^{H} r_h^{-1}\,w_h^2\,s_{h2}^2. \quad (7)$$

Thus, comparing (7) with (2), the bias of the jackknife variance estimator (6) is

$$\text{Bias}(\hat{V}_J) \doteq -E\left\{f_1 \sum_{h=1}^{H} (r_h^{-1} - n_h^{-1})\ s_{h2}^2\right\}.$$

Therefore, if the first-phase sampling rate is negligible in the sense of $f_1 \doteq 0$, the bias is negligible, *i.e.*, the bias = $o(n^{-1})$. Otherwise, the variance estimator underestimates the variance.

To consider a bias-corrected jackknife method, instead of (5), we consider

$$\bar{y}_{h2}^{(k)} = \begin{cases} (r_h - \delta_h)^{-1}(r_h \bar{y}_{h2} - \delta_h y_k) & \text{if } k \in A_{h2} \\ \bar{y}_{h2} & \text{if } k \notin A_{h2}, \end{cases} \quad (8)$$

where $\delta_h$ is to be determined. In (5), $\delta_h = 1$ was used. The jackknife variance estimator using (8) instead of (5) is asymptotically equivalent to

$$\hat{V}_J \doteq n^{-1}(1 - f_1) \sum_{h=1}^{H} w_h (\bar{y}_{h2} - \bar{y}_2)^2 + (1 - f_1) \sum_{h=1}^{H} \frac{(r_h - 1)\ \delta_h^2}{(r_h - \delta_h)^2}\ w_h^2\ s_{h2}^2.$$

Thus, the asymptotic bias is

$$\text{Bias}(\hat{V}_J) \doteq E\left[\sum_{h=1}^{H}\left\{(1 - f_1)\ \frac{(r_h - 1)\ \delta_h^2}{(r_h - \delta_h)^2} - \frac{1}{r_h}\left(1 - f_1 \frac{r_h}{n_h}\right)\right\} w_h^2\ s_{h2}^2\right].$$

The asymptotic bias is zero if

$$\delta_h = \frac{r_h}{1 + \sqrt{r_h (r_h - 1)}/d_h}$$

where $d_h = \sqrt{(1 - f_1 r_h n_h^{-1})/(1 - f_1)}$. Hence, with such determined $\delta_h$ in equation (8), the resulting jackknife variance estimator is approximately unbiased without assuming $f_1 \doteq 0$.

## 3. Proposed method

The proposed method in Section 2 is now extended to a more general first-phase sampling design. To do this, we need to assume that the replication variance estimator of the form

$$\hat{V}_1 = \sum_{k=1}^{L} c_k (\hat{\theta}^{(k)} - \hat{\theta})^2,$$

where $\hat{\theta} = \sum_{i \in A_1} w_i\ y_i$, and $\hat{\theta}^{(k)} = \sum_{i \in A_1} w_i^{(k)} y_i$, is consistent for the variance of $\hat{\theta}$ under the single (first) stage sampling design. That is,

$$\frac{\hat{V}_1}{\text{Var}(\hat{\theta})} - 1 = o_p(1). \quad (9)$$

Here $L$ is the number of replicates. For most of the measurable designs, which are designs with all positive joint inclusion probabilities, we can construct a replication variance estimator satisfying (9) even when the sample rate $f = n/N$ is large. For example, see Fay (1984) and Flyer (1987). Brick and Morganstein (1996) describes the basic algorithm for WesVar, a commercially available software for replication variance estimation in survey sampling.

In this section, we also consider a more challenging case of stratified unequal probability sampling for the second phase. More specifically, the second phase sampling considered is unequal probability Poisson sampling within the second-phase strata. Fuller (1998) also considered Poisson sampling in the second phase and argued that Poisson sampling in the second phase sampling is a good approximation. An example of this in the context of forest surveys is that, in addition to forest types, the photo-interpreters can also identify tree density and tree height from the aerial photos taken in the first phase, which can be used to construct the second phase selection probabilities within each stratum (forest type).

In this section, we will focus on the REE-type estimator first since it is more efficient than the DEE-type, and extension to the DEE is discussed in Section 4. Let $w_i$ be the first-phase sampling weight and let $w_{i2}$ be the inverse of the conditional probability in the second-phase. That is, $w_{i2} = \pi_{i2}^{-1}$ where $\pi_{i2} = \Pr(i \in A_{h2} \mid i \in A_{h1})$. The REE-type estimator can be written as

$$\bar{y}_{tp} = \frac{1}{N} \sum_{h=1}^{H} \hat{N}_{h1}\, \bar{y}_{h2} \quad (10)$$

where $\hat{N}_{h1} = \sum_{i \in A_{h1}} w_i$ and $\bar{y}_{h2} = (\sum_{i \in A_{h2}} w_i \pi_{i2}^{-1})^{-1} \sum_{i \in A_{h2}} w_i \pi_{i2}^{-1} y_i$. In KNF, $\pi_{i2}$ is assumed to be constant within the second-phase stratum.

We consider a replication-based approach for variance estimation of the REE-type estimator (10) when $\pi_{i2}$ is not necessarily constant within the second-phase stratum. We consider the special case when the second-phase sampling design is Poisson sampling. Using the replication method satisfying (9), the KNF-type variance estimator can be applied to estimate the variance of $\bar{y}_{tp}$ in this situation. That is,

$$\hat{V}_{\text{KNF}} = \sum_{k=1}^{L} c_k (\bar{y}_{tp}^{(k)} - \bar{y}_{tp})^2, \quad (11)$$

where

$$\bar{y}_{tp}^{(k)} = \frac{1}{N} \sum_{h=1}^{H} \hat{N}_{h1}^{(k)}\, \bar{y}_{h2}^{(k)} \quad (12)$$

with $\bar{y}_{h2}^{(k)} = (\sum_{i\in A_{h2}} w_i^{(k)}\pi_{i2}^{-1})^{-1}\sum_{i\in A_{h2}} w_i^{(k)}\pi_{i2}^{-1}y_i$ and $\hat{N}_{h1}^{(k)} = \sum_{i\in A_{h1}} w_i^{(k)}$, and $c_k$ is a factor associated with replicate $k$ determined by the replication method. Under Poisson sampling in the second phase, we have the following asymptotic bias:

$$\text{Bias}(\hat{V}_{\text{KNF}}) = -\frac{1}{N^2}\sum_{h=1}^{H}\sum_{i\in U_h}\pi_{i2}^{-1}(1-\pi_{i2})(y_i-\bar{Y}_h)^2, \quad (13)$$

where $U_h$ is the set of indices of population elements in stratum $h$ and $\bar{Y}_h = N_h^{-1}\sum_{i\in U_h} y_i$. A sketched proof of (13) is presented in Appendix A.

An asymptotically unbiased estimator of the bias (13) is

$$\hat{V}_{\text{bias}} = -\frac{1}{N^2}\sum_{h=1}^{H}\sum_{i\in A_{h2}} w_i\,\pi_{i2}^{-2}(1-\pi_{i2})(y_i-\bar{y}_{h2})^2. \quad (14)$$

The bias is negligible if $n/N \doteq 0$. Thus, we can safely ignore the bias of the KNF-type variance estimator when the first-phase sampling rate is negligible. The bias can be arbitrarily large if the first-phase sampling rate $n/N$ is not negligible. KNF also discuss a bias-correction replication method using additional replicates, which can lead to a large number of replicates. Creating additional replicates for bias-correction can be cumbersome for large scale surveys.

We consider an alternative bias-corrected replication variance estimator that does not require creating additional replicates. To develop a replication-based bias-corrected variance estimator, define a random variable

$$\delta_{ki} \overset{\text{indep}}{\sim} \text{Bernoulli}(p_k), \quad (15)$$

where $p_k$ is to be determined. Let

$$\hat{V}_{\text{KNF}}^{*} = \sum_{k=1}^{L} c_k(\bar{y}_{tp}^{*(k)} - \bar{y}_{tp})^2 \quad (16)$$

where

$$\bar{y}_{tp}^{*(k)} = \frac{1}{N}\sum_{h=1}^{H}\hat{N}_{h1}^{(k)}\,\bar{y}_{h2}^{*(k)} \quad (17)$$

with $\hat{N}_{h1}^{(k)} = \sum_{i\in A_{h1}} w_i^{(k)}$,

$$\bar{y}_{h2}^{*(k)} = \frac{\sum_{i\in A_{h2}} w_i^{(k)} M_{i2}^{(k)}\pi_{i2}^{-1}y_i}{\sum_{i\in A_{h2}} w_i^{(k)} M_{i2}^{(k)}\pi_{i2}^{-1}} \quad (18)$$

with

$$M_{i2}^{(k)} = 1 + (\delta_{ki} - p_k)b_i \quad (19)$$

and $b_i$ is also to be determined. By construction, $E_*(\delta_{ki} - p_k) = 0$, where $E_*$ denotes that the expectation is taken with respect to the mechanism in (15). Thus, the replicates (18) create additional variation in the replication weights, where the additional variation in (18) comes from the distribution (15). A suitable choice of $p_i$ and $b_i$ can make the resulting variance estimator consistent.

Under the regularity conditions discussed in KNF, we have

$$E_*(\hat{V}_{\text{KNF}}^{*}) = \hat{V}_{\text{KNF}} + N^{-2}\sum_{h=1}^{H}\sum_{i\in A_{h2}} w_i^2\,b_i^2\,\pi_{i2}^{-2}\,u\,(y_i-\bar{y}_{h2})^2 + o_p(n^{-1}), \quad (20)$$

where $u = \sum_{k=1}^{L} c_k p_k(1-p_k)$. A sketched proof of (20) is presented in Appendix B. If $b_i$ are determined by

$$b_i = \sqrt{(1-\pi_{i2})w_i^{-1}u^{-1}}, \quad (21)$$

the variance estimator (16) is consistent because the second term in (20) cancels out $\hat{V}_{\text{bias}}$ in (14). This is true even when the first-phase sampling rate $n/N$ is not negligible. To guarantee nonnegative replication weights in (18), we require that $b_i$ in (19) is $\le 1$. If we set $p_k = 0.5$, then

$$b_i = \sqrt{\frac{4\,(1-\pi_{i2})\,w_i^{-1}}{\sum_{k=1}^{L} c_k}},$$

which is less than or equal to 1 if $\sum_{k=1}^{L} c_k \ge 4$. In fact, the $p_k$'s can be chosen to be any number between 0 and 1 as long as the resulting $b_i$ in (21) is less than or equal to 1.

## 4. Extensions

In this section, we consider some extensions of the proposed replication method to types of two-phase estimators other than the REE in (10).

### 4.1 Double expansion estimator

In two-phase sampling, the double expansion estimator, termed by Kott and Stukel (1997), is also used. The double expansion estimator (DEE) has the simple form

$$\bar{y}_{\text{DEE}} = \frac{1}{N}\sum_{h=1}^{H}\sum_{i\in A_{h2}} w_i\,\pi_{i2}^{-1}\,y_i. \quad (22)$$

When the second-phase sample is a stratified random sample, $\pi_{i2} = r_h/n_h$ and the KNF method can be applied using the replicate

$$\bar{y}_{\text{DEE}}^{(k)} = \frac{1}{N}\sum_{h=1}^{H}\left(\frac{\sum_{i\in A_{h1}} w_i^{(k)}\,w_i^{-1}}{\sum_{i\in A_{h2}} w_i^{(k)}\,w_i^{-1}}\right)\sum_{i\in A_{h2}} w_i^{(k)}\,y_i.$$

The KNF variance estimator for DEE is consistent when the first-phase sampling rate is negligible. When the first-phase sampling rate is not negligible, we can use the replication method proposed in Section 3. The proposed replication method for the DEE creates replicates,

$$\bar{y}_{\text{DEE}}^{*(k)} = \frac{1}{N}\sum_{h=1}^{H}\sum_{i\in A_{h2}} w_i^{(k)} w_{i2}^{*(k)} y_i, \tag{23}$$

where

$$w_{i2}^{*(k)} = M_{i2}^{(k)} \frac{\sum_{i\in A_{h1}} w_i^{(k)} w_i^{-1}}{\sum_{i\in A_{h2}} w_i^{(k)} w_i^{-1} M_{i2}^{(k)}},$$

and $M_{i2}^{(k)}$ is the replication factor defined in (19). The bias of the replication variance estimator using replicate (23) is negligible if the replicates are constructed to satisfy (21).

If the second-phase sample is an unequal probability sample within each stratum, the replication method such as (23) is not directly applicable. The DEE in (22) is generally less efficient than the REE in (10). Note that the REE in (10) can also be expressed as

$$\bar{y}_{\text{REE}} = \frac{1}{N}\sum_{h=1}^{H}\sum_{i\in A_{h2}} w_i w_{i2}^{*} y_i, \tag{24}$$

where

$$w_{i2}^{*} = \pi_{i2}^{-1} \frac{\sum_{i\in A_{h1}} w_i}{\sum_{i\in A_{h2}} w_i \pi_{i2}^{-1}}. \tag{25}$$

The replicates (17) can be written

$$\bar{y}_{\text{REE}}^{*(k)} = \frac{1}{N}\sum_{h=1}^{H}\sum_{i\in A_{h2}} w_i^{(k)} w_{i2}^{*(k)} y_i, \tag{26}$$

where

$$w_{i2}^{*(k)} = M_{i2}^{(k)}\pi_{i2}^{-1} \frac{\sum_{i\in A_{h1}} w_i^{(k)}}{\sum_{i\in A_{h2}} w_i^{(k)} M_{i2}^{(k)} \pi_{i2}^{-1}} \tag{27}$$

and $M_{i2}^{(k)}$ is defined in (19).

### 4.2 Regression estimator

In two-phase sampling, auxiliary variables that are observed in the first-phase sample can be further used at the estimation stage. The two-phase regression estimator of the population total can be written in the form

$$\hat{Y}_{t,\text{REG}} = \hat{\mathbf{T}}_{x,1}' \hat{\boldsymbol{\beta}}_2 \tag{28}$$

where $\hat{\mathbf{T}}_{x,1} = \sum_{i\in A_1} w_i \mathbf{x}_i$ is the vector of estimated population totals of the control variable $\mathbf{x}_i$ estimated with the first-phase sample and $\hat{\boldsymbol{\beta}}_2 = (\sum_{i\in A_2} w_i w_{i2}^{*} \mathbf{x}_i \mathbf{x}_i')^{-1} \sum_{i\in A_2} w_i w_{i2}^{*} \mathbf{x}_i y_i$ is a vector of estimated regression coefficients estimated with the second-phase sample and $w_{i2}^{*}$ is given by (25). Note that the regression estimator in (28) can incorporate the stratified sampling design in the second-phase if $\mathbf{x}_i$ includes the vector of stratum indicators.

Using the arguments of Section 3, the $k^{\text{th}}$ replicate for $\hat{Y}_{t,\text{REG}}$ can be constructed by

$$\hat{Y}_{t,\text{REG}}^{(k)} = \hat{\mathbf{T}}_{x,1}^{(k)\prime} \hat{\boldsymbol{\beta}}_2^{(k)}, \tag{29}$$

where

$$\hat{\mathbf{T}}_{x,1}^{(k)} = \sum_{i\in A_1} w_i^{(k)} \mathbf{x}_i$$

$$\hat{\boldsymbol{\beta}}_2^{(k)} = \left(\sum_{i\in A_2} w_i^{(k)} w_{i2}^{*(k)} \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \sum_{i\in A_2} w_i^{(k)} w_{i2}^{*(k)} \mathbf{x}_i y_i$$

and $w_{i2}^{*(k)}$ is defined in (27).

The replication method (29) can be directly applicable to the two-phase calibration estimator that was discussed in Hidiroglou and Särndal (1998). If $H = 1$, then the replicate of $\hat{\boldsymbol{\beta}}_2$ in (29) reduces to

$$\hat{\boldsymbol{\beta}}_2^{(k)} = \left(\sum_{i\in A_2} w_i^{(k)} M_{i2}^{(k)} \pi_{i2}^{-1} \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \sum_{i\in A_2} w_i^{(k)} M_{i2}^{(k)} \pi_{i2}^{-1} \mathbf{x}_i y_i.$$

## 5. Simulation study

To study the finite sample performance of the proposed estimators, we conducted a limited simulation study. In the simulation, we first generated an artificial finite population of size $N = 1{,}000$ with five variables $(z_i, q_i, x_i, y_i, u_i)$, where the population elements are independently generated from $z_i \sim \exp(1) + 2$; $q_i \sim \chi^2(1) + 2$; $x_i \sim N(2, 1)$; $u_i \sim \text{Unif}\{1, 2, 3, 4\}$, where $\text{Unif}\{1, ..., G\}$ denotes a discrete uniform distribution with support $\{1, ..., G\}$; and

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 q_i + e_i$$

with $(\beta_0, \beta_1, \beta_2, \beta_3) = (0, 2, 1, 1)$ and $e_i \sim N(0, 1)$. The variables $z_i$, $q_i$, $x_i$, $u_i$, and $e_i$ are mutually independent. The stratum for the second-phase sampling was defined using variable $u_i$. Variable $x_i$ was used to compute the two-phase regression estimator (28) with $\mathbf{x}_i = (1, x_i)'$, variable $z_i$ was used as a size measure for the unequal probability sampling in the first phase sampling, and variable $q_i$ was used as a size measure for the unequal probability sampling in the second phase sampling.

To obtain unequal probability samples for this simulation study, we used either Poisson sampling or Rao-Sampford sampling (Rao 1965 and Sampford 1967), with selection probabilities proportional to the measure of the size variable. Note that the final sample size is random under Poisson sampling but is fixed under Rao-Sampford sampling.

The simulation setup employed a $2 \times 3 \times 2$ factorial structure with three factors. The factors are

1. Sampling for the first-phase sample (2): Simple random sampling of size $n = 200$ versus the Rao-Sampford sampling of size $n = 200$ using $z_i$ as the measure of size.

2. Sampling for the second-phase sample (3): Stratified random sampling of size $r_h = 25$, stratified Poisson sampling with expected sample size $r_h = 25$ using $q_i$ as the size measure for the unequal probability sampling, and stratified Rao-Sampford sampling of size $r_h = 25$ using $q_i$ as the size measure for the unequal probability sampling.
3. Variance estimation methods (2): The KNF estimator (11) without additional replication versus the proposed variance estimator using (16) were computed based on the jackknife method.

From the finite population generated above, we generated $B = 5{,}000$ independent Monte Carlo samples for simulation. For the designs with Rao-Sampford sampling in the first phase, we used the jackknife variance estimation method proposed by Berger (2007), which gives a consistent estimator of the first phase sampling variance. The parameter of interest is the population mean of the $y$ variable. From each Monte Carlo sample, we computed two point estimators, the REE in (24) and the regression estimator (REG) in (28) using the auxiliary variable $(1, x_i)$. Relative biases of the variance estimators were computed by dividing the Monte Carlo bias of the variance estimator by the Monte Carlo variance of the point estimator.

Table 1 shows the mean and variance of the two point estimators. For point estimation, the regression estimator is significantly more efficient than the REE for this population because the auxiliary variable $x$ is correlated with the study variable $y$. The theoretical asymptotic variance of the regression estimator under simple random sampling in the first phase and stratified random sampling in the second phase is approximately equal to

$$\left(\frac{1}{200}-\frac{1}{1{,}000}\right)8+\left(\frac{1}{100}-\frac{1}{200}\right)4 = 0.052$$

and the theoretical asymptotic variance of the REE under the same design is, approximately, $(1/100-1/1{,}000)\,8 = 0.072$, which is consistent with the numerical results in Table 1. The Rao-Sampford sampling in the second phase is slightly more efficient than the Poisson sampling because of the fixed sample size in the Rao-Sampford sampling.

Table 2 shows the relative bias (RB) and coefficient of variation (CV) of the two variance estimators. Relative biases of the variance estimators were computed by dividing the Monte Carlo bias of the variance estimator by the Monte Carlo variance of the point estimator. Coefficients of variation of the variance estimator were computed by dividing the Monte Carlo standard error of the variance estimator by the Monte Carlo average of the variance estimator.

**Table 1**
**Mean and variance of the point estimators (5,000 samples)**

| Estimator | First-phase Sampling | Second-Phase Sampling | Mean | Variance |
|---|---|---|---|---|
| REE | SRS | St. SRS | 10.0 | 0.0749 |
| | | St. Poi | 10.0 | 0.0784 |
| | | St. RS | 10.0 | 0.0754 |
| | RS | St. SRS | 10.0 | 0.0768 |
| | | St. Poi | 10.0 | 0.0827 |
| | | St. RS | 10.0 | 0.0781 |
| REG | SRS | St. SRS | 10.0 | 0.0540 |
| | | St. Poi | 10.0 | 0.0510 |
| | | St. RS | 10.0 | 0.0495 |
| | RS | St. SRS | 10.0 | 0.0551 |
| | | St. Poi | 10.0 | 0.0531 |
| | | St. RS | 10.0 | 0.0515 |

REE: reweighted expansion estimator (23),
REG: regression estimator (27),
SRS: Simple random sampling,
RS: Rao-Sampford sampling,
St. SRS: Stratified simple random sampling,
St. Poi: Stratified Poisson sampling,
St. RS: Stratified Rao-Sampford sampling.

**Table 2**
**Relative bias (RB) and coefficient of variation (CV) for the variance estimators (5,000 samples)**

| Method | Estimator | First-phase Sampling | Second-Phase Sampling | RB (%) | CV(%) |
|---|---|---|---|---|---|
| KNF | REE | SRS | St. SRS | -11.25 | 18.22 |
| | | | St. Poi | -9.56 | 18.67 |
| | | | St. RS | -7.75 | 15.35 |
| | | RS | St. SRS | -8.05 | 18.61 |
| | | | St. Poi | -9.03 | 20.84 |
| | | | St. RS | -5.73 | 17.27 |
| | REG | SRS | St. SRS | -6.76 | 22.32 |
| | | | St. Poi | -6.06 | 15.81 |
| | | | St. RS | -3.26 | 12.82 |
| | | RS | St. SRS | -4.17 | 21.74 |
| | | | St. Poi | -3.64 | 16.92 |
| | | | St. RS | -3.20 | 13.78 |
| New | REE | SRS | St. SRS | 0.09 | 18.23 |
| | | | St. Poi | -1.23 | 19.70 |
| | | | St. RS | -0.04 | 16.06 |
| | | RS | St. SRS | 0.78 | 19.78 |
| | | | St. Poi | -2.07 | 21.26 |
| | | | St. RS | 1.00 | 17.67 |
| | REG | SRS | St. SRS | -0.61 | 22.00 |
| | | | St. Poi | -0.57 | 16.55 |
| | | | St. RS | -0.08 | 13.36 |
| | | RS | St. SRS | 0.67 | 22.86 |
| | | | St. Poi | -0.01 | 16.97 |
| | | | St. RS | 0.59 | 14.02 |

KNF: Kim *et al.* (2006) variance estimator without additional replicates for bias correction,
New: the proposed variance estimator (16),
REE: reweighted expansion estimator (23),
REG: regression estimator (27),
SRS: Simple random sampling,
RS: Rao-Sampford sampling,
St. SRS: Stratified simple random sampling,
St. Poi: Stratified Poisson sampling,
St. RS: Stratified Rao-Sampford sampling.

In this simulation, because the first-phase sampling fraction is not negligible $(n/N = 0.2)$, the KNF variance estimator without additional replicates underestimates the true variance and the proposed variance estimator estimates the variance with smaller bias, less than 3% in absolute values in all cases, which is consistent with the theory in Section 3 and Section 4. The absolute value of the relative biases in the KNF variance estimator are big because, although in (29) the variance due to $\hat{\mathbf{T}}_{x,1}$ is consistently estimated, the variance due to $\hat{\boldsymbol{\beta}}_2$ is underestimated without additional replicates. The relative biases in our proposed variance estimator are reduced because replicates (18) create additional variation in the replication weights through additional perturbation $\delta_k$ drawn from a properly chosen distribution. The proposed variance estimator shows slightly bigger CVs than the KNF method because it involves extra randomness due to generating $\delta_{ki}$ from (15).

## 6. Concluding remarks

Replication variance estimation under two-phase sampling is an importance practical problem in survey sampling and the KNF method is a useful tool in this direction. In this article, we propose an extension of the KNF method in that it can be directly applicable when the first-phase sampling rate is non-negligible, without increasing the number of replicates. The proposed method is also applicable to unequal probability Poisson sampling within each stratum in the second-phase sample. Although the theory has been developed only under Poisson sampling in the second phase, the simulation results in section 5 show that the proposed method works reasonably well for other unequal probability sampling designs, such as the Rao-Sampford sampling design. Since the proposed replication method provides consistent variance estimators for population means, it can be readily applied to other finite population parameters which are smooth functions of population means.

In some large scale surveys, the number of replicates can be quite large because it uses the same number of replicates for the first-phase sample. If one wishes to reduce the number of replicates further, the method of Fuller (1998) or Kim and Sitter (2003) can be considered. Further investigation in this direction will be a topic of future study.

## Acknowledgements

The research was supported by a Cooperative Agreement No. 68-3A75-4-122 between the USDA Natural Resources Conservation Service and the Center for Survey Statistics and Methodology at Iowa State University. The authors wish to thank Wayne Fuller and two anonymous referees for helpful comments.

## Appendix

### A. Proof of (13)

Let $\mathbf{a} = (a_1, ..., a_N)$ where $a_i$ is the extended version of the second-phase sampling indicator as discussed in Kim *et al.* (2006). That is, $a_i = 1$ if unit $i$ is selected for the second-phase sample once it is in the first-phase sample and $a_i = 0$ otherwise.

By assumption (9), conditional on $\mathbf{a}$, we have

$$\sum_{k=1}^{L} c_k (\bar{y}_{h2}^{(k)} - \bar{y}_{h2})^2 = \mathrm{Var}(\bar{y}_{h2} \mid \mathbf{a}) + o_p(n^{-1}).$$

Thus, the bias of $\sum_{k=1}^{L} c_k (\bar{y}_{h2}^{(k)} - \bar{y}_{h2})^2$ as an estimator for $\mathrm{Var}(\bar{y}_{h2})$ is then equal to, ignoring $o(n^{-1})$ terms,

$$E\{\mathrm{Var}(\bar{y}_{h2} \mid \mathbf{a})\} - \mathrm{Var}(\bar{y}_{h2}) = \mathrm{Var}\{E(\bar{y}_{h2} \mid \mathbf{a})\}.$$

Using the extended definition of $a_i$, we have

$$E(\bar{y}_{h2} \mid \mathbf{a}) = \frac{\sum_{i \in U_h} \pi_{i2}^{-1} a_i y_i}{\sum_{i \in U_h} \pi_{i2}^{-2} a_i}$$

and, by the Poisson sampling assumption of $a_i$'s,

$$\mathrm{Var}\left( \frac{\sum_{i \in U_h} \pi_{i2}^{-1} a_i y_i}{\sum_{i \in U_h} \pi_{i2}^{-1} a_i} \right) = N_h^{-2} \sum_{i \in U_h} \pi_{i2}^{-1} (1 - \pi_{i2}) (y_i - \bar{Y}_h)^2 + o(N^{-1}). \quad \text{(A.1)}$$

Thus, the bias of the KNF variance estimator is of the form (13) under the Poisson sampling assumption of $a_i$.

### B. Proof of (20)

For each $k$,

$$\bar{y}_{tp}^{*(k)} - \bar{y}_{tp} = \bar{y}_{tp}^{*(k)} - \bar{y}_{tp}^{(k)} + \bar{y}_{tp}^{(k)} - \bar{y}_{tp},$$

where $\bar{y}_{tp}^{(k)}$ is defined in (12). Thus,

$$\hat{V}_{\mathrm{KNF}}^{*} = \sum_{k=1}^{L} c_k (\bar{y}_{tp}^{*(k)} - \bar{y}_{tp})^2 = \sum_{k=1}^{L} c_k (\bar{y}_{tp}^{(k)} - \bar{y}_{tp})^2 + 2 \sum_{k=1}^{L} c_k (\bar{y}_{tp}^{(k)} - \bar{y}_{tp})(\bar{y}_{tp}^{*(k)} - \bar{y}_{tp}^{(k)}) + \sum_{k=1}^{L} c_k (\bar{y}_{tp}^{*(k)} - \bar{y}_{tp}^{(k)})^2. \quad \text{(B.1)}$$

By the construction of $\bar{y}_{tp}^{*(k)}$, we have

$$E_*(\bar{y}_{tp}^{*(k)}) = \bar{y}_{tp}^{(k)} + o_p(n^{-1}). \quad \text{(B.2)}$$

Also, writing $q_{ki} = M_{i2}^{(k)} - 1$, we have $q_{ki} = O_p(n^{-1/2})$ and we can apply a Taylor expansion to get

$$\bar{y}_{h2}^{*(k)} = \bar{y}_{h2}^{(k)} + \frac{\sum_{i\in A_{h2}} w_i^{(k)} \pi_{i2}^{-1} q_{ki} (y_i - \bar{y}_{h2}^{(k)})}{\sum_{i\in A_{h2}} w_i^{(k)} \pi_{i2}^{-1}} + o_p(n^{-1}). \quad \text{(B.3)}$$

Also, because

$$\frac{1}{N_h} \sum_{i\in A_{h2}} w_i^{(k)} \pi_{i2}^{-1} z_i - \frac{1}{N_h} \sum_{i\in A_{h2}} w_i \pi_{i2}^{-1} z_i = O_p(n^{-1})$$

for any $z$ variable with bounded fourth moments, it can be shown that (B.3) reduces to

$$\bar{y}_{h2}^{*(k)} = \bar{y}_{h2}^{(k)} + \frac{\sum_{i\in A_{h2}} w_i \pi_{i2}^{-1} q_{ki} (y_i - \bar{y}_{h2})}{\sum_{i\in A_{h2}} w_i \pi_{i2}^{-1}} + o_p(n^{-1}).$$

Hence, we can write

$$\sum_{k=1}^{L} c_k (\bar{y}_{tp}^{*(k)} - \bar{y}_{tp}^{(k)})^2 =$$

$$\sum_{k=1}^{L} c_k \left\{ N^{-1} \sum_{h=1}^{H} \sum_{i\in A_{h2}} w_i \pi_{i2}^{-1} q_{ki} (y_i - \bar{y}_{h2}) \right\}^2 + o_p(n^{-1}). \quad \text{(B.4)}$$

Inserting (B.2) and (B.4) into (B.1), we have

$$E_*(\hat{V}_{\text{KNF}}^*) = \hat{V}_{\text{KNF}} + \frac{1}{N^2} \sum_{k=1}^{L} c_k \sum_{h=1}^{H} \sum_{i\in A_{h2}} w_i^2 E_*(q_{ki}^2) \pi_{i2}^{-2} (y_i - \bar{y}_{h2})^2 + o_p(n^{-1}),$$

and because $E_*(q_{ki}^2) = p_k(1 - p_k)\, b_i^2$, we have (20).

## References


Berger, Y.G. (2007). A jackknife variance estimator for unistage stratified samples with unequal probabilities. *Biometrika*, 94, 953-964.

Binder, D.A., Babyak, C., Brodeur, M., Hidiroglou, M. and Jocelyn, W. (2000). Variance estimation for two-phase stratified sampling. *The Canadian Journal of Statistics*, 28, 751-764.

Breidt, F.J. and Fuller, W.A. (1993). Regression weighting for multipurpose samplings. *Sankhyā*, B, 55, 297-309.

Brick, J.M., and Morganstein, D. (1996). WesVarPC: Software for computing variance estimates from complex designs. *Proceedings of the 1996 Annual Research Conference*, U.S. Bureau of the Census, 861-866.

Cochran, W.G. (1977). *Sampling Techniques*. New York: John Wiley & Sons, Inc.

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

Flyer, P. (1987). Finite population correction for replication estimates of variance. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 732-736.

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

Fuller, W.A. (2003). Estimation for multiple phase samples. In *Analysis of Survey Data*, (Eds., R.L. Chambers and C.J. Skinner). Wiley, Chichester, England, 307-322.

Hidiroglou, M.A. (2001). Double sampling. *Survey Methodology*, 27, 143-154.

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

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

Kim, J.K., and Sitter, R.R. (2003). Efficient variance estimation for two-phase sampling. *Statistica Sinica*, 13, 641-653.

Kott, P.S., and Stukel, D.M. (1997). Can the jackknife be used with a two-phase Sample? *Survey Methodology*, 23, 81-89.

Lu, W., and Sitter, R.R. (2006). Disclosure risk and variance estimation. *Proceedings of Statistics Canada international symposium series*, 11-522-XIE.

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

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. (1973). On double sampling for stratification and analytical surveys. *Biometrika*, 60, 125-133.

Rao, J.N.K., and Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79, 811-822.

Rao, J.N.K., and Sitter, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82, 453-460.

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 probability of selection. *Biometrika*, 54, 499-513.

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

Survey Methodology, June 2011
Vol. 37, No. 1, pp. 75-94
Statistics Canada, Catalogue No. 12-001-X



# Cost efficiency of repeated cluster surveys

**Stanislav Kolenikov and Gustavo Angeles** [1]

Abstract

We analyze the statistical and economic efficiency of different designs of cluster surveys collected in two consecutive time periods, or waves. In an independent design, two cluster samples in two waves are taken independently from one another. In a cluster-panel design, the same clusters are used in both waves, but samples within clusters are taken independently in two time periods. In an observation-panel design, both clusters and observations are retained from one wave of data collection to another. By assuming a simple population structure, we derive design variances and costs of the surveys conducted according to these designs. We first consider a situation in which the interest lies in estimation of the change in the population mean between two time periods, and derive the optimal sample allocations for the three designs of interest. We then propose the utility maximization framework borrowed from microeconomics to illustrate a possible approach to the choice of the design that strives to optimize several variances simultaneously. Incorporating the contemporaneous means and their variances tends to shift the preferences from observation-panel towards simpler panel-cluster and independent designs if the panel mode of data collection is too expensive. We present numeric illustrations demonstrating how a survey designer may want to choose the efficient design given the population parameters and data collection cost.

Key Words: Longitudinal study; Cluster samples; DHS; NHIS.

## 1. Introduction

To analyze the dynamics of social, behavioral or population health phenomena, researchers and policymakers need to obtain information on characteristics of the population on multiple occasions. Complex design surveys are the most frequently used sources of information for large populations, such as a country as a whole. Besides the standard considerations in single-shot surveys, *e.g.*, stratification and clustering, other issues may be important in surveys collected over two or more time periods. In such surveys, the total cost and the total survey error are affected by an overlap among consecutive samples, (informative) sample attrition, time-in-sample or conditioning effects, and other dynamic factors.

For the purposes of estimation of change from repeated surveys, it is often desirable to have high temporal correlation of the observation units which can be achieved by administering the survey to the same sampling and/or observation units. In longitudinal surveys, the same observation units (individuals, households) are revisited for several periods, potentially indefinitely many periods (the US Panel Study of Income Dynamics (PSID), British Household Panel Study (BHPS) and others). A compendium of information on the longitudinal studies can be found at the Institute for Social and Economics Research web site, http://iser.essex.ac.uk/ulsc/keeptrack/index.php). In rotating panel surveys, the observation units are recruited into the sample for a few periods, then rotated out of the sample, and surveyed again at a later time. Examples of rotating panel surveys include the US Current Population Survey (CPS) (Binder and Hidiroglou 1988, Eckler 1955, Rao and Graham 1964) and a number of environmental surveys (Fuller 1999, McDonald 2003, Scott 1998). Yet another option is to use the same primary sampling units (PSUs) in different waves, but sample the observation units (secondary sampling units, SSUs) independently. Surveys collected in this way include international Demographic and Health Surveys (DHS) and the US National Health Interview Survey (NHIS).

We shall concentrate on surveys collected in two time periods, or waves, using a two-stage cluster design in each wave of data collection. We consider three possible designs differing in the amount and depth of overlap of sampling units over time. The sample designer can simply ignore any possible effects arising from the sample overlap, and take two independent samples in two periods of time. We shall refer to this design as the *independent* design. Alternatively, the sample designer may find it beneficial to recycle the PSUs from one wave to another. If the designer finds it difficult to track the SSUs from one wave to another, the subsamples within clusters can be taken independently in two waves of data collection. We shall refer to this design as the *cluster-panel* design. If an utmost precision is essential, the fully longitudinal design will attempt to locate all individuals who responded in the first wave, and solicit the second interview. To distinguish this design from the cluster-panel design, we shall refer to it as the *observation-panel* design.

1. Stanislav Kolenikov, Department of Statistics, 146 Middlebush Hall, University of Missouri, Columbia, MO 65211-6100, U.S.A. E-mail: kolenikovs@missouri.edu; Gustavo Angeles, Associate Director of the Center for Evaluation Research, National Institute of Public Health, Mexico, Mexico. E-mail: gangeles@insp.mx.

A particular aspect that we found important in survey management, but underaddressed in the existing literature, is the implementation cost (Groves 1989). The traditional cost models such as those used in derivation of Neyman-Tchuprow optimal allocation design (Neyman 1938) can be extended to include terms related to the cost of the first visit to the cluster and ultimate observation unit, as well as the cost of consecutive visits. The cost of revisiting the cluster is likely to be lower on the second occasion. There is no need to create new maps and set up frames. The same interviewers can be used to conduct interviews in subsequent waves of data collection. Cooperation with community leaders has been established earlier, if it is important, as it is in some traditional societies. The effect of the panel mode of data collection at the individual level is less clear. If the household that was interviewed in earlier waves moved out and would have to be located, possibly in different geographic area, the (average) cost of the panel interview goes up. The likelihood of such circumstances increases with longer intervals between surveys typical for the developing countries surveys: the intervals between waves of DHS are usually about 5-7 years. On the other hand, if a less expensive interview mode can be used after the first round, (*e.g.*, a phone interview instead of the personal visit), the cost of the panel interview goes down.

This paper brings together statistical and economic considerations in the choice of the appropriate design and its parameters. We assume the survey designer can be interested in estimating the change in the population mean between two time periods, and/or the means themselves. We introduce a sketchy population in Section 2, and compute the design variances of the means and their differences for the three sampling designs of our interest.

To incorporate economic aspects of data collection, we introduce a relatively simple cost model for a repeated cluster survey in Section 3. We set up and solve optimization problems to obtain the optimal sample sizes for the three considered designs. By plugging in the estimates of the statistical parameters (variances and autocorrelations) and cost components (cluster-level and individual-level costs), the survey designer can compare the numeric values of the variances to choose the best design. Section 4 illustrates this approach and shows that each of the designs may be the best one, depending on the parameter values. The intuitive results (*e.g.*, the higher cost of data collection and lower autocorrelations of the observed characteristics make panel modes of data collection less appealing) are given an analytic justification and quantitative backing.

While Sections 2-4 deal with the efficiency in estimating the difference in means only, more realistic goals of data collection efforts would include contemporaneous characteristics and their variances. To this end, Section 5 introduces a utility maximization framework describing the survey designer's choice of the sampling scheme. This framework provides an aggregated objective function that combines several design criteria. The results are again as expected: if the more expensive panel modes of data collection result in smaller sample sizes, the estimates of the means are less efficient than in simpler designs. The only way to justify these efficiency losses is by a drastic improvement in the estimation of the difference that can only occur with higher autocorrelations. Such effects are also illustrated in Section 5. Section 7 concludes. Proofs are given in the Appendix.

## 2. Design variances

Let the population consist of $N$ clusters, or PSUs, in both time periods, and each cluster consist of $M$ individuals, or SSUs. Out of these, an SRS of $1 < n_t \le N$ clusters is taken at time $t = 1, 2,$ and an SRS of $1 < m_t \le M$ individuals is taken in each cluster that is present in the sample at time $t$. Let the index $i$ denote PSUs, and the index $j$, SSUs. Thus the typical measurement will be denoted as $Y_{tij}$ in the population, and $y_{tij}$ in the sample. The population totals $T[\cdot]$ and their estimates $t[\cdot]$ can then be found as follows:

cluster total:

$$T_{ti\cdot}[Y] = \sum_{j=1}^{M} Y_{tij},\ t_{ti\cdot}[y] = \frac{M}{m}\sum_{j=1}^{M} y_{tij},$$

population total:

$$T_{t\cdot\cdot}[Y] = \sum_{i=1}^{N} Y_{ti\cdot},\ t_{t\cdot\cdot}[y] = \frac{N}{n}\sum_{i=1}^{N} t_{ti\cdot}[y]. \qquad (2.1)$$

The means per observation units are

$$\bar{Y}_{ti\cdot} = \frac{1}{M}\sum_{i=1}^{M} Y_{tij} = \frac{T_{ti\cdot}[Y]}{T_{ti\cdot}[1]},\ \bar{y}_{ti\cdot} = \frac{1}{m}\sum_{i=1}^{m} y_{tij} = \frac{t_{ti\cdot}[y]}{t_{ti\cdot}[1]},$$

$$\bar{Y}_{t\cdot\cdot} = \frac{T_{t\cdot\cdot}[Y]}{T_{t\cdot\cdot}[1]} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} Y_{tij}}{NM},\ \bar{y}_{t\cdot\cdot} = \frac{t_{t\cdot\cdot}[y]}{t_{t\cdot\cdot}[1]} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} y_{tij}}{nm}. \qquad (2.2)$$

The variance of $Y$ and its within- and between-cluster components are

$$S_t^2 = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}(Y_{tij} - \bar{Y}_{t\cdot\cdot})^2}{N\bar{M} - 1}, \qquad (2.3)$$

$$S_{twi}^2 = \frac{\sum_{j=1}^{M}(Y_{tij} - \bar{Y}_{ti\cdot})^2}{M - 1},\ \bar{S}_{tw}^2 = \frac{1}{N}\sum_{i=1}^{N} S_{twi}^2, \qquad (2.4)$$

$$S_{tb}^2 = \frac{\sum_{i=1}^{N}(\overline{Y}_{ti\cdot} - \overline{Y}_{t\cdot\cdot})^2}{N-1}. \quad (2.5)$$

The characteristic of primary interest is the change in the means,

$$D = \overline{Y}_{2\cdot\cdot} - \overline{Y}_{1\cdot\cdot}, \quad (2.6)$$

estimated by

$$d = \overline{y}_{2\cdot\cdot} - \overline{y}_{1\cdot\cdot}. \quad (2.7)$$

An attractive property of this estimator for analysts and data users is its internal consistency: the estimator of the difference is the difference of the estimators. If the samples in consecutive periods overlap only partially, then composite or GLS estimators (Fuller 1999, Hansen, Hurwitz and Madow 1953, Patterson 1950, Rao and Graham 1964, Wolter 2007) have better efficiency.

In what follows, we assume all sampling procedures to be simple random sampling without replacement. For the contemporaneous mean, the variance is given by (Cochran 1977, Th. 10.1):

$$\mathrm{V}[\overline{y}_{t\cdot\cdot}] = \left(1 - \frac{n}{N}\right)\frac{S_{tb}^2}{n} + \left(1 - \frac{m}{M}\right)\frac{\overline{S}_{tw}^2}{nm}. \quad (2.8)$$

For simplicity and clarity of exposition, we shall often be making an assumption of symmetric conditions:

$$S_{1wi}^2 = S_{2wi}^2 = S_{wi}^2,\ \overline{S}_{1w}^2 = \overline{S}_{2w}^2 = \overline{S}_w^2,\ S_{1b}^2 = S_{2b}^2 = S_b^2. \quad (2.9)$$

Analytic derivations are possible without these assumptions, but become extremely cumbersome. Besides, it is unrealistic to think that the survey designer could know the characteristics of the future population. Thus (2.9) should be viewed as a reasonable working model.

### 2.1 Independent design

*Proposition* 1. *Let $n_1$ out of $N$ clusters and $m_1$ out of $M$ observation units in selected clusters be taken without replacement at time $t = 1$. Let $n_2$ out of $N$ clusters and $m_2$ out of $M$ observation units in selected clusters be taken without replacement at time $t = 2$, with sampling performed independently from that at time $t = 1$. Then*

$$\mathrm{V}_\iota(d) = \left(1 - \frac{n_1}{N}\right)\frac{S_{1b}^2}{n_1} + \left(1 - \frac{n_2}{N}\right)\frac{S_{2b}^2}{n_2} + \left(1 - \frac{m_1}{M}\right)\frac{\overline{S}_{1w}^2}{n_1 m_1} + \left(1 - \frac{m_2}{M}\right)\frac{\overline{S}_{2w}^2}{n_2 m_2}. \quad (2.10)$$

The result follows immediately from (2.8) by independence of the two samples. The subindex of the variance $\iota$ stands for the "independent design". Under the symmetric conditions of (2.9), if the sample sizes are the same in two periods, $n_1 = n_2 = n$ and $m_1 = m_2 = m$, then

$$\mathrm{V}_{e,\iota}[d] = 2\left(1 - \frac{n}{N}\right)\frac{S_b^2}{n} + 2\left(1 - \frac{m}{M}\right)\frac{\overline{S}_w^2}{nm}, \quad (2.11)$$

where the subindex $e, \iota$ stands for "equal variances, independent design".

### 2.2 Cluster-panel design

*Proposition* 2. *Let $n$ out of $N$ clusters be sampled without replacement in the first period and be used in both time periods. Let $m$ out of $M$ observation units be sampled without replacement independently in two periods. Then*

$$\mathrm{V}_c[d] = \left(1 - \frac{n}{N}\right)\frac{S_{1b}^2 + S_{2b}^2 - 2\rho^{\mathrm{I}} S_{1b} S_{2b}}{n} + \left(1 - \frac{m}{M}\right)\frac{\overline{S}_{1w}^2 + \overline{S}_{2w}^2}{nm},$$

$$\rho^{\mathrm{I}} = \frac{1}{S_{1b} S_{2b}(N-1)}\sum_{i=1}^{N}(\overline{Y}_{1i\cdot} - \overline{Y}_{1\cdot\cdot})(\overline{Y}_{2i\cdot} - \overline{Y}_{2\cdot\cdot}). \quad (2.12)$$

Here, subindex $c$ stands for the "cluster-panel design", and $\rho^{\mathrm{I}}$ is the intertemporal correlation, or autocorrelation, of the cluster means. The superscript I denotes the first stage of sampling. If $\rho^{\mathrm{I}}$ is positive, then the cluster-panel design is more efficient than the independent design for fixed values of $n$ and $m$. Under the symmetry conditions,

$$\mathrm{V}_{e,c}[d] = 2\left(1 - \frac{n}{N}\right)\frac{S_b^2(1-\rho^{\mathrm{I}})}{n} + 2\left(1 - \frac{m}{M}\right)\frac{\overline{S}_w^2}{mn}, \quad (2.13)$$

where the subindex $e, c$ stands for the "equal variances, cluster-panel design".

### 2.3 Observation-panel design

*Proposition* 3. *Let $n$ out of $N$ clusters and $m$ out of $M$ observation units be sampled without replacement in the first period and be used in both time periods. Then*

$$\mathrm{V}_o[d] = \left(1 - \frac{n}{N}\right)\frac{S_{1b}^2 + S_{2b}^2 - 2\rho^{\mathrm{I}} S_{1b} S_{2b}}{n} + \left(1 - \frac{m}{M}\right)\frac{\overline{S}_{1w}^2 + \overline{S}_{2w}^2 - 2\rho^{\mathrm{II}}\overline{S}_{1w}\overline{S}_{2w}}{nm},$$

$$\rho^{\mathrm{II}} = \frac{1}{\overline{S}_{1w}\overline{S}_{2w}N(M-1)}\sum_{i=1}^{N}\sum_{j=1}^{M}(Y_{1ij} - \overline{Y}_{1i\cdot})(Y_{2ij} - \overline{Y}_{2i\cdot}). \quad (2.14)$$

Subindex $o$ stands for the "observation-panel design". Under the assumption of symmetric conditions,

$$V_{e,o}[d] = 2\left(1-\frac{n}{N}\right)\frac{(1-\rho^{I})S_b^2}{n} + 2\left(1-\frac{m}{M}\right)\frac{(1-\rho^{II})\bar{S}_w^2}{nm},$$

$$\rho^{II} = \frac{1}{\bar{S}_w^2 N(M-1)}\sum_{i=1}^{N}\sum_{j=1}^{M}(Y_{1ij}-\bar{Y}_{1i\cdot})(Y_{2ij}-\bar{Y}_{2i\cdot}) \quad (2.15)$$

with corresponding $e,o$ subindex for the "equal variances, observation-panel design".

Here, $\rho^{II}$ is the intertemporal correlation, or autocorrelation, of the individual observations within clusters. The superscript II stands for the second stage of sampling. If $\rho^{II}$ is positive, then the observation-panel design is more efficient than the cluster-panel design for fixed values of $n$ and $m$.

How are the two autocorrelations that appear in (2.15) related? Conceptually, one can think of any number of possible relations between them. Let us introduce a superpopulation model

$$Y_{tij} = \mu_t + a_{ti} + \varepsilon_{tij}, \quad E_\xi[a_{ti}] = 0, \quad E_\xi[\varepsilon_{tij}] = 0, \quad (2.16)$$

in which $a_{ti}$ and $\varepsilon_{sij}$ are independent of one another for all $s, t = 1, 2$. The subindex $\xi$ stands for the superpopulation model expectations. The case of $\rho^{I} = 0$ and $\rho^{II} = 1$ occurs when the changes in the cluster means occur independently between clusters ($E_\xi[a_{1i}a_{2i}] = 0$), but the individuals retain their positions within the cluster, $\varepsilon_{1ij} = \varepsilon_{2ij}$. The case of $\rho^{I} = 1$ and $\rho^{II} = 0$ occurs when the cluster random effects are the same in both periods, $a_{1i} = a_{2i}$, while the individual random effects are uncorrelated ($E_\xi[\varepsilon_{1ij}\varepsilon_{2ij}] = 0$). Neither of these situations is entirely realistic. However, it can probably be expected that the individual, rather than the cluster, dynamics are a more important source of variation over time, thus making the relations $\rho^{II} \geq \rho^{I} \geq 0$ the most plausible ones. We shall study in numeric examples of Sections 4 and 5 the extent to which the choice of the best design is sensitive to the relation between the two correlations.

## 3. Costs for repeated cluster samples

In this section we shall analyze the cost efficiency of cluster samples when one wants to estimate the difference between two sample means from two different periods.

Some discussion of the costs of cluster sampling is given in Kish (1995, Section 8.3B), Thompson (1992, Section 12.5), and Lehtonen and Pahkinen (2004). More mathematical details are available in Hansen *et al.* (1953, volume II, Section 6.11), with the variance formulas corrected for finite populations.

### 3.1 Notation and cost models

Let us assume the following cost structure, which is an extension of Kish (1995) for repeated surveys:

- $c_1^{I}$ is the cluster level cost at time $t = 1$ for clusters that are used *in the first wave only*;
- $c_2^{I}$ is the cluster level cost for a *new* cluster at time $t = 2$;
- $c_{12}^{I}$ is the cluster level cost for clusters in which the data are collected in both periods $t = 1$ and $t = 2$ (PSU panel cost);
- $c_1^{II}$ is the individual level cost at time $t = 1$ for individuals that are observed *in the first wave only*;
- $c_2^{II}$ is the individual level cost at time $t = 2$ for individuals that are observed *in the second wave only*;
- $c_{12}^{II}$ is the individual level cost if the unit is observed in both periods in the observation-panel design (SSU panel cost);
- $C_0$ is the total budget allocated to the field work in both time periods.

Roman superscripts denote the sampling stage. Arabic subscripts correspond to the occasion at which the sample is taken. The cluster level costs include the cost of sampling the clusters, obtaining the PSU maps, collecting community data, local interviewer training, *etc.* The individual level costs are mostly those of the personal interviews with the ultimate observation units. The total cost $C_0$ is thought of as the variable cost of the survey that is directly related to the number of sampled units. Fixed cost, such as the cost of preparing the survey instrument and other organization-level costs are not part of $C_0$.

### 3.2 Independent design

The budget constraint for the independent design is given by

$$C_0 = c_1^{I} n_1 + c_1^{II} n_1 m_1 + c_2^{I} n_2 + c_2^{II} n_2 m_2. \quad (3.1)$$

The first two terms are the costs of the first wave of data collection, and the last two terms, of the second wave.

*Proposition* 4. *If the survey setting parameters are the same in the two time periods*:

$$c_1^{I} = c_2^{I} = c^{I}, \; c_1^{II} = c_2^{II} = c^{II}, \quad (3.2)$$

*then the optimal sample sizes and the resulting variances are given by*

$$m = \sqrt{\frac{c^{\mathrm{I}}}{c^{\mathrm{II}}}\frac{\overline{S}_w^2}{S_b^2 - \overline{S}_w^2/M}},$$

$$n = \frac{C_0}{2\{c^{\mathrm{I}} + [c^{\mathrm{I}}c^{\mathrm{II}}\overline{S}_w^2/(S_b^2 - \overline{S}_w^2/M)]^{1/2}\}},$$

$$\mathrm{V}_{e,\iota}[d] = \frac{4\left[c^{\mathrm{I}} + \sqrt{c^{\mathrm{I}}c^{\mathrm{II}}\overline{S}_w^2/(S_b^2 - \overline{S}_w^2/M)}\right]}{C_0} \times \left[S_b^2 + \left(\sqrt{\frac{c^{\mathrm{II}}}{c^{\mathrm{I}}}\frac{S_b^2 - \overline{S}_w^2/M}{\overline{S}_w^2}} - \frac{1}{M}\right)\overline{S}_w^2\right] - \frac{2}{N}S_b^2. \quad (3.3)$$

In equations (3.3), the sample sizes $n$ and $m$ are treated as continuous variables. In practice, the nearest integer should be used, with a minimum of 2 necessary to estimate the appropriate variance component, and the maxima of $N$ and $M$, respectively.

The number of observations sampled within a cluster depends only on the relative costs at the cluster and the observation level, $c^{\mathrm{I}}/c^{\mathrm{II}}$, and relative variances $S_b^2/\overline{S}_w^2$, or equivalently the intraclass correlation. Greater interview cost $c^{\mathrm{II}}$ prevents the sample designer from using more observations: an increase in $c^{\mathrm{II}}$ leads to a decrease in both $m$ and $n$. Greater cluster-level cost leads to redistribution of the sampled units: $n$ decreases with $c^{\mathrm{I}}$, while $m$ increases with it. Greater within-cluster variance $\overline{S}_w^2$ necessitates a greater number of observations $m$ to be taken within a cluster to maintain overall precision. Greater between-cluster variance $S_b^2$ necessitates a greater number of clusters $n$ to be sampled. Finally, the total survey budget $C_0$ affects the number of clusters $n$, but not the subsample size $m$. As a result, the variance of $d$ is inversely proportional to $C_0$.

The non-symmetric situation can be treated as a by-product of the first order conditions derived in the proof (see Appendix). However, no analytic solution is available in that case.

### 3.3 Cluster-panel design

The budget constraint for the cluster-panel design is given by

$$C_0 = c_{12}^{\mathrm{I}}n + c_1^{\mathrm{II}}nm_1 + c_2^{\mathrm{II}}nm_2. \quad (3.4)$$

The first term is the cluster-level cost associated with the sample design, and the remaining two terms are the costs of collecting individual-level data in the first and the second waves, respectively.

*Proposition* 5. *The sample sizes for the cluster-panel design are given by*

$$m_1 = 2C_0\Big/c_1^{\mathrm{II}}\left(1 + \frac{\overline{S}_{2w}^2}{\kappa\overline{S}_{1w}^2} + \sqrt{D}\right),$$

$$m_2 = \kappa m_1,$$

$$n = \frac{C_0}{c_{12}^{\mathrm{I}} + c_1^{\mathrm{II}}m_1 + c_2^{\mathrm{II}}m_2},$$

$$\kappa = \sqrt{\frac{c_1^{\mathrm{II}}}{c_2^{\mathrm{II}}}\frac{\overline{S}_{2w}^2}{\overline{S}_{1w}^2}}, \quad (3.5)$$

*provided that*

$$D = \left(1 + \frac{\overline{S}_{2w}^2}{\kappa\overline{S}_{1w}^2}\right)^2 + 8\frac{(1-\rho^{\mathrm{I}})S_b^2C_0}{\overline{S}_{1w}^2c_1^{\mathrm{II}}} - 4\frac{C_0}{c_1^{\mathrm{II}}M}\left(1 + \frac{\overline{S}_{2w}^2}{\kappa\overline{S}_{1w}^2}\right) \geq 0.$$

The variance of the difference estimator is found by plugging these expressions into (2.13). Under the assumptions of symmetric conditions in two rounds of the survey (2.9) and (3.2),

$$D = 4 - 8\frac{C_0}{Mc^{\mathrm{II}}} + 8\frac{(1-\rho^{\mathrm{I}})S_b^2C_0}{\overline{S}_w^2c^{\mathrm{II}}},$$

$$m_1 = m_2 = m = \frac{C_0}{c^{\mathrm{II}} + \sqrt{(c^{\mathrm{II}})^2 - \frac{2c^{\mathrm{II}}C_0}{M} + \frac{2(1-\rho^{\mathrm{I}})S_b^2C_0c^{\mathrm{II}}}{\overline{S}_w^2}}},$$

$$n = \frac{C_0}{c_{12}^{\mathrm{I}} + 2c^{\mathrm{II}}m} = \frac{C_0}{c_{12}^{\mathrm{I}} + 2C_0\Big/\left[1 + \sqrt{1 - \frac{2c^{\mathrm{II}}C_0}{Mc^{\mathrm{II}}} + \frac{2(1-\rho^{\mathrm{I}})S_b^2C_0}{\overline{S}_w^2c^{\mathrm{II}}}}\right]},$$

and $\mathrm{V}_{e,c}[d]$ can be found from (2.13).

Interestingly, the number of the SSUs depends on the SSU costs $c^{\mathrm{II}}$, but not on the PSU costs $c_{12}^{\mathrm{I}}$. An increase in the intracluster correlation, or increase in $S_b^2$, or decrease in $S_w^2$, predictably leads to decrease in the optimal number of SSUs and increase in the optimal number of PSUs. The dependence of the design parameters on the survey budget $C_0$ is non-trivial. For very small surveys, the number of units per cluster is proportional to $C_0$, and the number of clusters is not affected by $C_0$. Indeed, if the characteristic demonstrates strong correlation between time periods, it would be preferable to get accurate estimates of the cluster means, and good accuracy of the overall difference estimator will follow. To put it differently, the first term in (2.13) is relatively small by virtue of the positive correlation coefficient $\rho^{\mathrm{I}}$, and the second term is inversely proportional

to $C_0$. For large surveys, $D \propto C_0$, so both the number of units per cluster and the number of clusters are proportional to $\sqrt{C_0}$. The first term in (2.13) is then inversely proportional to $\sqrt{C_0}$, and the second term is inversely proportional to $C_0$. An increase in the budget of the survey will affect all terms, although to a different extent.

### 3.4 Observation-panel design

The budget constraint for the observation-panel design is given by

$$C_0 = c_{12}^{\mathrm{I}} n + c_{12}^{\mathrm{II}} nm. \tag{3.6}$$

The first term is the cluster-level cost, and the second term is the cost of individual interviews.

*Proposition* 6. *The optimal sample sizes for the observation-panel design are given by*

$$m = \sqrt{\frac{c_{12}^{\mathrm{I}}}{c_{12}^{\mathrm{II}}} \frac{(1-\rho^{\mathrm{II}})\bar{S}_w^2}{(1-\rho^{\mathrm{I}})S_b^2 - (1-\rho^{\mathrm{II}})\bar{S}_w^2 / M}},$$
$$n = \frac{C_0}{c_{12}^{\mathrm{I}} + \sqrt{\frac{(1-\rho^{\mathrm{II}})\bar{S}_w^2 c_{12}^{\mathrm{I}} c_{12}^{\mathrm{II}}}{(1-\rho^{\mathrm{I}})S_b^2 - (1-\rho^{\mathrm{II}})\bar{S}_w^2 / M}}}. \tag{3.7}$$

*The design variance of the resulting difference estimator is*

$$\begin{aligned} \mathrm{V}_{e,o}[d] = \frac{2}{C_0}\Bigg\{ & (1-\rho^{\mathrm{I}}) S_b^2 c_{12}^{\mathrm{I}} \\ & + (1-\rho^{\mathrm{II}})\bar{S}_w^2 \sqrt{c_{12}^{\mathrm{I}} c_{12}^{\mathrm{II}} \frac{(1-\rho^{\mathrm{I}})S_b^2 - (1-\rho^{\mathrm{II}})\bar{S}_w^2 / M}{(1-\rho^{\mathrm{II}})\bar{S}_w^2}} \\ & + \left[ (1-\rho^{\mathrm{I}}) S_b^2 - \frac{1}{M}(1-\rho^{\mathrm{II}})\bar{S}_w^2 \right] \\ & \times \sqrt{\frac{(1-\rho^{\mathrm{II}})\bar{S}_w^2 c_{12}^{\mathrm{I}} c_{12}^{\mathrm{II}}}{(1-\rho^{\mathrm{I}})S_b^2 - (1-\rho^{\mathrm{II}})\bar{S}_w^2 / M}} \\ & + (1-\rho^{\mathrm{II}})\bar{S}_w^2 \left( c_{12}^{\mathrm{II}} - \frac{c_{12}^{\mathrm{I}}}{M} \right) \Bigg\} - \frac{2(1-\rho^{\mathrm{I}})S_b^2}{N}. \end{aligned} \tag{3.8}$$

The sample size expressions (3.7) resemble the ones for the independent design, equation (3.3), with the cost of data collection in a single wave replaced by the cost of panel data collection, and the variance components $S_b^2$ and $\bar{S}_w^2$ replaced by $(1-\rho^{\mathrm{I}})S_b^2$ and $(1-\rho^{\mathrm{II}})\bar{S}_w^2$. The second stage sampling size $m$ only depends on the relative cost at the cluster and observation levels, and on the ratio of the variance components augmented by the autocorrelations. Hence, like in the independent design, the dependency of the sample size on the scale of the survey is only through $n \propto C_0$, and the variance of the difference decreases inversely proportional to $C_0$.

Extending the relations between the functional forms of equations (3.3) and (3.8), we can establish the general relations between the two designs:

*Proposition* 7. *If* $M \gg 1$ *and* $N \gg 1$, *then* $\mathrm{V}_{e,\iota}[d] \gtrless \mathrm{V}_{e,o}[d]$ *if*

$$2\left(\sqrt{c^{\mathrm{I}} S_b^2} + \sqrt{c^{\mathrm{II}} \bar{S}_w^2}\right)^2 \gtrless \left[\sqrt{c_{12}^{\mathrm{I}}(1-\rho^{\mathrm{I}})S_b^2} + \sqrt{c_{12}^{\mathrm{II}}(1-\rho^{\mathrm{II}})\bar{S}_w^2}\right]^2. \tag{3.9}$$

Unfortunately, the variance for the cluster-panel design that can be obtained by combining the results of Proposition 5 with (2.13), does not permit an equally lucid comparison.

## 4. Numeric illustration

To illustrate how the characteristics of population (variances and autocorrelations) and the data collection process (costs) affect the choice of the most efficient design, we consider a numeric example. Let us choose the basic setup with symmetric conditions, and let the parameter values be:

$$\begin{aligned} & N = 10{,}000, \quad M = 1{,}000, \quad S_b = 100, \\ & S_w = 400, \quad \rho^{\mathrm{I}} = 0.1, \quad \rho^{\mathrm{II}} = 0.35, \\ & c_1^{\mathrm{II}} = c_2^{\mathrm{II}} = 1, \quad c_{12}^{\mathrm{II}} = 3, \quad c_1^{\mathrm{I}} = c_2^{\mathrm{I}} = 10, \\ & c_{12}^{\mathrm{I}} = 18, \quad C_0 = 20{,}000. \end{aligned} \tag{4.1}$$

The cost structure implies that the cost of collecting the initial information for a cluster is the cost of ten interviews, while the cost of the followup in the same cluster is only eight interviews. On the other hand, getting the second interview with the same unit is twice as expensive as getting the first interview.

With these parameters, the sample sizes and design variances are:

$$\begin{aligned} & m_{e,\iota} = 12, \quad m_{e,c} = 12, \quad m_{e,o} = 8, \\ & n_{e,\iota} = 455, \quad n_{e,c} = 476, \quad n_{e,o} = 476, \\ & m_{e,\iota} n_{e,\iota} = 5{,}460, \quad m_{e,c} n_{e,c} = 5{,}712, \quad m_{e,o} n_{e,o} = 3{,}808, \\ & \mathrm{V}_{e,\iota}[d] = 99.86, \quad \mathrm{V}_{e,c}[d] = 91.37, \quad \mathrm{V}_{e,o}[d] = 90.20. \end{aligned} \tag{4.2}$$

The observation-panel design is 1.2% more efficient than the cluster-panel design, and 10.7% more efficient than the independent design. However, it has a notably smaller total sample size, only $2/3$ of the cluster-panel design sample size and 70% of the independent design sample size.

Of course these finding are highly specific to the parameters of the population and the cost structure. Can we describe general patterns of how the variances, and hence the relative efficiency of different designs, change with those parameters? The variances in (4.2) are derived from 13 parameters given in (4.1), and it is difficult to make meaningful statements about all of these parameters simultaneously. Below, we shall attempt to provide two-dimensional cross-sections of this 13-dimensional space and give graphical illustrations of the variability of the design variances, and hence the domains of optimality of each design, as we vary two parameters at a time. We provide the graphs of variances of the designs involved (typically, the cluster-panel design with dotted lines, the observation-panel design with dashed lines, and the independent design with dash-dotted lines. For most plots, the independent design is not affected by the variations of the parameters that make up the axis of the plots, and hence omitted). We also show the relative efficiency of different designs, marking the domains of the parameter space in yellow/light gray if the independent design is the most efficient one; in green/medium gray if the cluster-panel design is the most efficient one; and in purple/dark gray if the observation-panel design is the most efficient one (R code used to produce graphs is available at http://web.missouri.edu/~kolenikovs/SMJ2011/).

Figure 1 shows how the design variances, and hence the most efficient design, vary with the panel costs of the PSU and SSU, $c_{12}^{\mathrm{I}}$ and $c_{12}^{\mathrm{II}}$. Obviously, these variations do not affect the variance of the independent design, which serves as a benchmark. Also, the variations in $c_{12}^{\mathrm{II}}$ do not affect the performance of the cluster-panel design, which corresponds to the dotted vertical iso-variance lines on the left panel. The dashed downward sloping lines are the iso-variance lines for the observation-panel design. Note that the lower left corner of the graph corresponds to the free lunch situation in which the second wave of data collection does not cost anything: the panel costs are equal to the single period cost, $c_{12}^{\mathrm{I}} = c_{1}^{\mathrm{I}}$, $c_{12}^{\mathrm{II}} = c_{1}^{\mathrm{II}}$. When the costs of the panel data collection are prohibitively high (the upper right corner of the graph), the independent design is the most efficient one. The point where all three designs have the same variances is $c_{12}^{\mathrm{I}} = 22$, $c_{12}^{\mathrm{II}} = 3.05$; *i.e.*, the cost of the second interview is 2.05 higher than the cost of the first interview, and the cluster-level costs in the second wave are 20% higher than in the first wave. Still, a positive autocorrelation justifies the reduction in the sample size of the observation-panel design as compared to the independent design. If the cluster level panel cost is lower and the second interview cost is higher, the cluster-panel design is the most efficient. For inexpensive second interviews, the most efficient design is the observation-panel design. The latter domain includes our baseline case with $c_{12}^{\mathrm{I}} = 18$ and $c^{\mathrm{II}} = 3$.

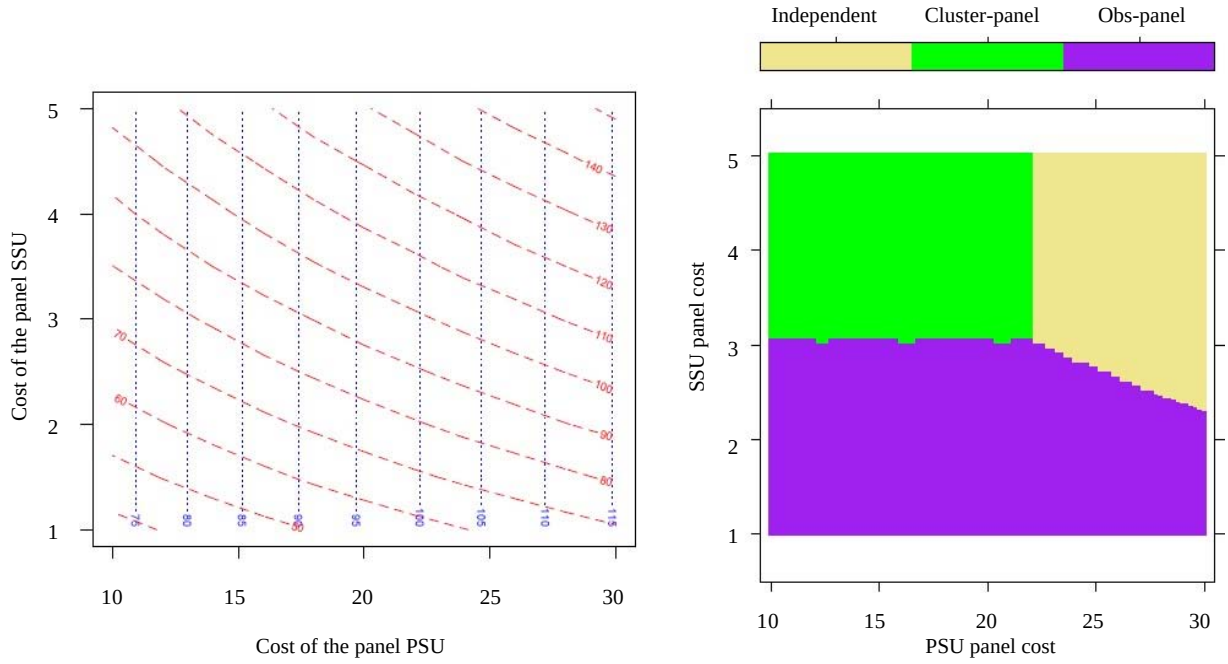


**Figure 1 Design variances as functions of the data collection costs $c_{12}^{\mathrm{I}}, c_{12}^{\mathrm{II}}$. Left: contour lines of $\mathrm{V}_{e,c}[d]$ (dotted) and $\mathrm{V}_{e,o}[d]$ (long dashed); $\mathrm{V}_{e,\iota} = 99.86$; right: domains of optimality of the three designs**

Figure 2 shows the changes in design variances associated with the changes in the autocorrelations $\rho^{I}$, $\rho^{II}$. The independent design variance is unaffected by these variations, and the cluster-panel design is unaffected by variations in $\rho^{II}$. The observation-panel design is more efficient for higher SSU autocorrelation, $\rho^{II} > 0.34$. Otherwise, the cluster-panel design provides lower variance.

Figure 3 investigates the impact of the cluster-level cost and autocorrelation on the choice of the design. The combinations of expensive second wave of data collection and low PSU autocorrelation in the upper left corner of the plot makes the independent design the most appealing one. Otherwise, the observation-panel design is the best one to use. Note that the contour lines for the cluster-panel and observation-panel designs are very close to one another, and differences in variances between the two designs are less than 2% in the whole parameter space of this plot.

Figure 4 investigates the impact of the observation-level cost and autocorrelation on the choice of the design. Neither the independent design nor the cluster-panel design variances are affected by variation of the parameters shown on this plot. The independent design variance is 99.86, while the cluster-panel design variance is 91.37, so the observation-panel design is compared to the latter only. High autocorrelations ($\rho^{II} \geq 0.6$) can justify very high cost of the second interview (up to fourfold compared to the first interview), but in the upper left corner of the plot corresponding to the low autocorrelations and high panel cost, the cluster-panel design performs better.

Figure 5 relates the design variances to the cluster-level costs of the survey. The horizontal axis is the cost in the first period, $c_1^I$, and the vertical axis is the additional cost of in the second period when the data are collected in a panel mode, $c_{12}^I - c_1^I$. The vertical axis is ignored for the independent design, as this parameter does not appear in the independent design. Also, by virtue of (4.1), $c_1^I = c_2^I$. The observation-panel design is uniformly better than the cluster-panel design for all parameter combinations on this graph, although the difference in variances does not exceed 2%. In the upper left corner, the additional cost of the panel mode of data collection is prohibitively high, and the independent design offers better performance.

Figure 6 shows the dependence of the most efficient design on the total budget of the survey and the cost of panel mode of data collection at the cluster level. For $C_0 > 10{,}000$, the observation-panel design performs better if $c_{12}^I < 22.7$, *i.e.*, if the additional cost of the panel mode of data collection at the cluster level does not exceed 127% of the initial cluster-level cost in the first wave. Interestingly, for some isolated parameter configurations in small surveys, the cluster-panel design can perform better than the observation-panel design that dominates the rest of the plot. The difference in design variances between the cluster-panel and observation-panel designs is less than 4% across all parameter combinations on this graph.

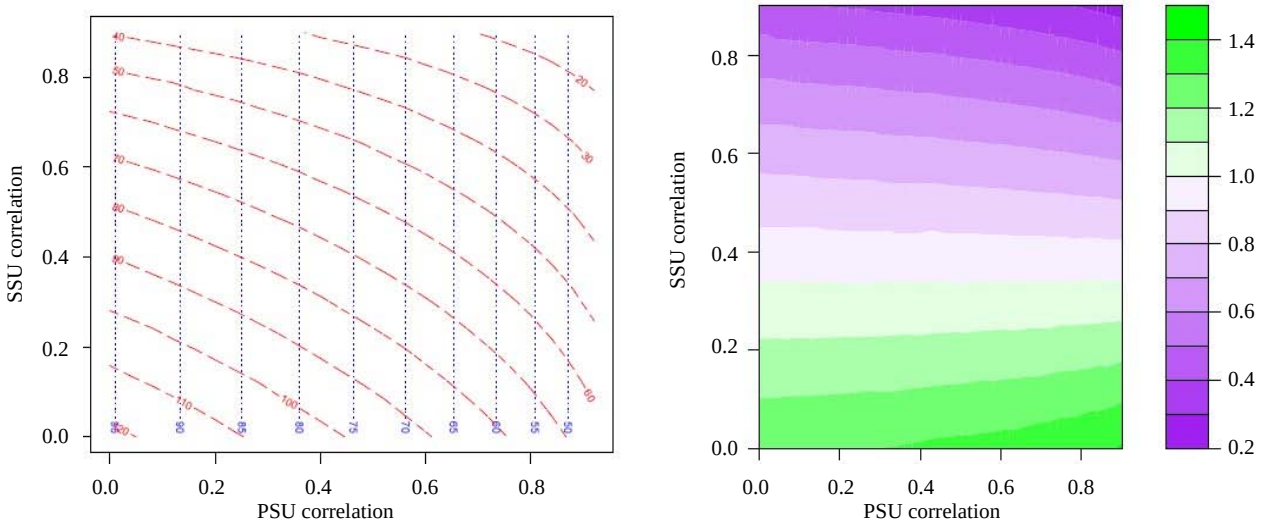


**Figure 2 Design variances as functions of the population correlations $\rho^{I}$, $\rho^{II}$. Left: contour lines of $V_{e,c}[d]$ (dotted) and $V_{e,o}[d]$ (long dashed); $V_{e,\iota} = 99.86$; right: ratio $V_{e,o}[d] / V_{e,c}[d]$**

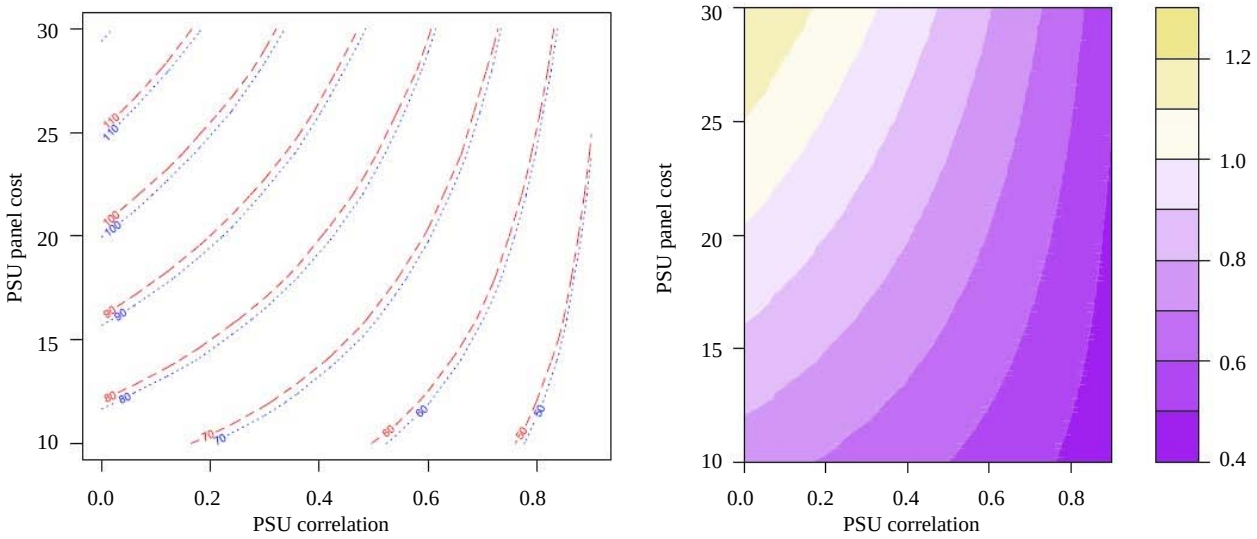


**Figure 3** Design variances as functions of the cluster-level autocorrelation $\rho^{I}$ and cost $c_{12}^{I}$. Left: contour lines of $V_{e,c}[d]$ (dotted) and $V_{e,o}[d]$ (long dashed); $V_{e,\iota} = 99.86$; right: ratio $V_{e,o}[d] / V_{e,\iota}[d]$

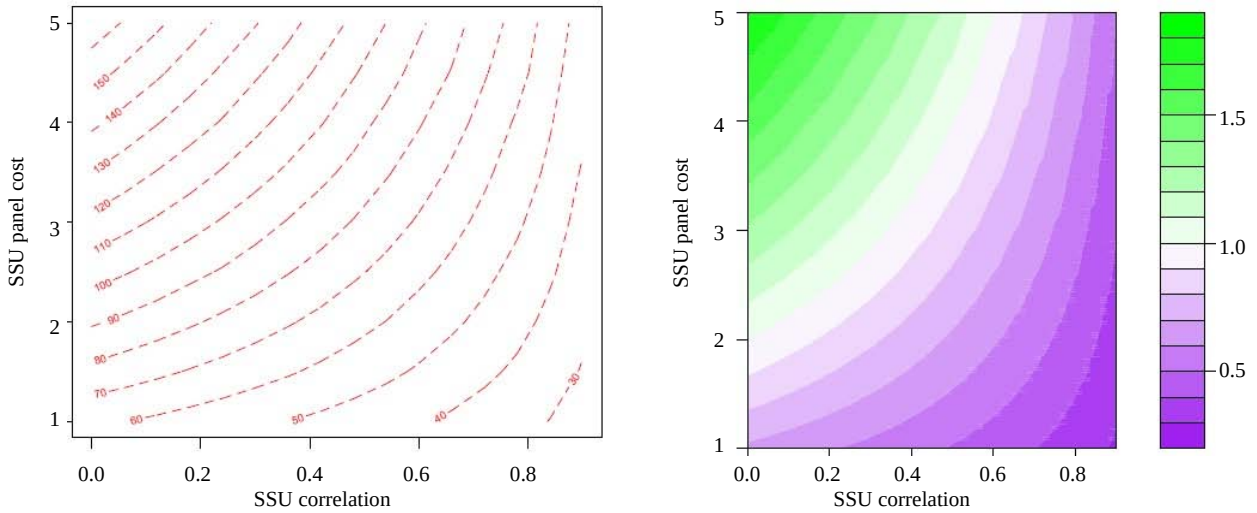


**Figure 4** Design variances as functions of the observation-level autocorrelation $\rho^{II}$ and cost $c_{12}^{II}$. Left: contour lines of $V_{e,o}[d]$ (long dashed); $V_{e,\iota} = 99.86$; $V_{e,c}[d] = 91.37$; right: ratio $V_{e,o}[d] / V_{e,c}[d]$

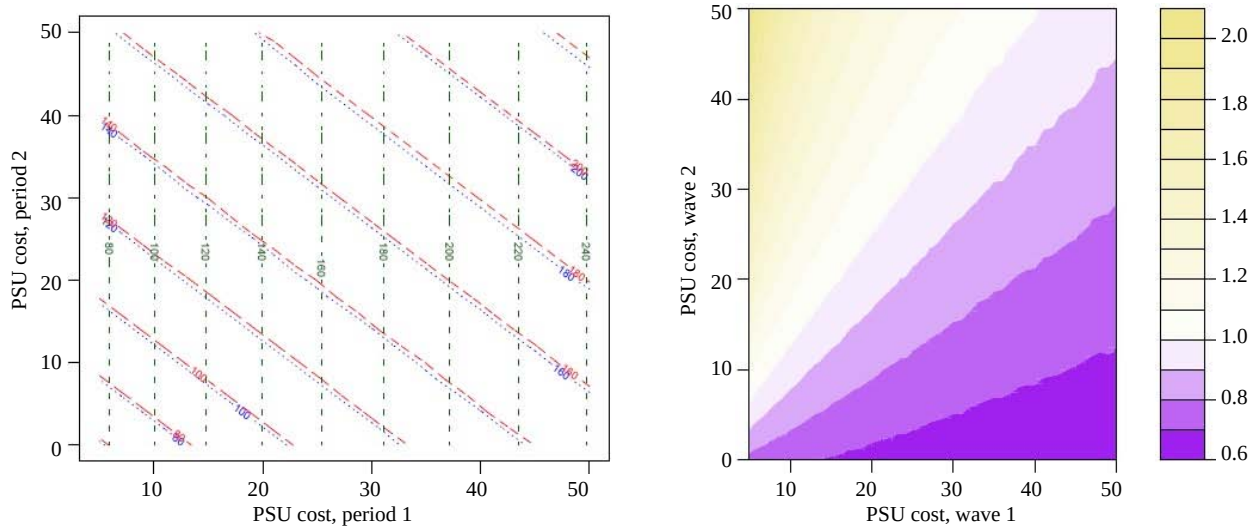


**Figure 5** Design variances as functions of the cluster level costs in the first wave, $c_1^{I}$, and in the second wave, $c_{12}^{I} - c_1^{I}$. Left: contour lines of $V_{e,c}[d]$ (dotted), $V_{e,o}[d]$ (long dashed) and $V_{e,\iota}[d]$ (dash-dotted); right: ratio $V_{e,o}[d] / V_{e,\iota}[d]$

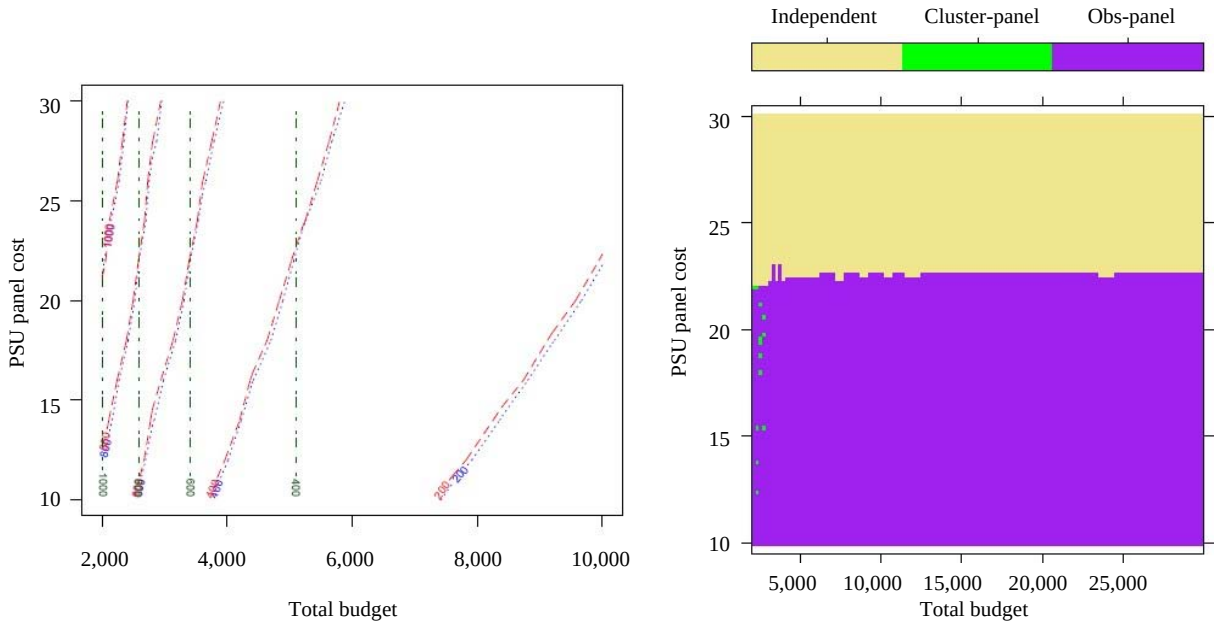


**Figure 6 Design variances as functions of the total budget $C_0$ and the PSU panel cost $c_{12}^{II}$. Left: contour lines of $V_{e,c}[d]$ (dotted), $V_{e,o}[d]$ (long dashed) and $V_{e,\iota}[d]$ (dash-dotted); right: domains of optimality of the three designs**

Overall, this numeric illustration shows that depending on the parameters of the population and costs of data collection, each of the three designs can be the most efficient one. Low correlations and high costs in the second wave tend to favor the independent design. Given that the initial six population parameters and five cost parameters may not be representative of many repeated surveys, a sensitivity analysis like the one performed here may be needed for any particular survey a statistician needs to design.

## 5. Survey design with multiple criteria

So far, our analysis was confined to estimation of the difference between the means in two waves of data collection of a single variable. Most large scale surveys are collected to study several characteristics, and to many users, the contemporaneous estimates are also of interest. To accommodate accuracy requirements associated with these different variables and different estimates, the survey designer must have several variances in mind when choosing the design to be implemented. This is a multicriterial optimization problem, and no single design will work best for all possible estimation problems. In the current context, the observation-panel design may give good estimates of the change when both PSU and SSU autocorrelations are high, but it may result in a small sample size if both PSUs and SSUs are expensive to follow up. Greater precision of the estimates for any single period could be obtained by switching to the cluster-panel or even independent designs.

Comparing different designs in this situation is possible with the standard microeconomic argument of utility maximization under budget constraints (Mas-Colell, Whinston and Green 1995). In the survey design context, the utility of the survey designer increases with the precision of the survey estimates, or equivalently decreases with survey variances. A simple functional form is given by Cobb-Douglas utility function:

$$U(\text{design}) = \mathrm{V}_{\text{design}}^{-\alpha_1}[\bar{y}_{1\cdot}]\,\mathrm{V}_{\text{design}}^{-\alpha_2}[\bar{y}_{2\cdot}]\,\mathrm{V}_{\text{design}}^{-\alpha_3}[d]. \quad (5.1)$$

Here, $\alpha_1$, $\alpha_2$ and $\alpha_3$ are positive constants describing the relative weights of the three design variances in decision-making process. Variances $\mathrm{V}[\bar{y}_1]$ and $\mathrm{V}[\bar{y}_2]$ in (5.2) are the variances of the means in cluster surveys given by (2.8). The variance of the difference estimator is (2.10), (2.12) or (2.14), depending on the design. The survey designer problem is then to maximize (5.1) subject to design-specific budget constraints (3.1), (3.4) or (3.6). Maximization is performed over the design parameters (mode of data collection, number of clusters in each time period, number of observations in each time period), given the characteristics of population (variances and autocorrelations) and the data collection process (costs).

Let us assume that the precision of each of the three estimates $\bar{y}_1$, $\bar{y}_2$ and $d$ is equally important to the decision maker, so $\alpha_1 = \alpha_2 = \alpha_3$. To have an objective function that is measured in the variance units and is on the same scale as variances, it will be convenient to define a multicriterial variance

$$V_{design} = (V_{design}[\bar{y}_{1\cdot\cdot}]\, V_{design}[\bar{y}_{2\cdot\cdot}]\, V_{design}[d])^{1/3}, \quad (5.2)$$

and express the optimization problem as minimization of this expression.

Analytic characterization of the design that optimizes (5.2) becomes quite cumbersome. Instead, we utilize a numeric illustration of the previous section to demonstrate how accounting for other design objectives affects the choice of the design. We should expect that for the designs with more expensive follow-ups ($c_{12}^{I} \geq c_1^{I} + c_2^{I}$, $c_{12}^{II} \geq c_1^{II} + c_2^{II}$), the simpler designs would be selected more often: the cluster-panel design may be preferred to the observation-panel design, and the independent design may be preferred to the cluster-panel design. For the baseline settings (4.1), we have

$$V_{e,\iota}[\bar{y}] = 49.93,\ V_{e,c}[\bar{y}] = 47.68,\ V_{e,o}[\bar{y}] = 61.69,$$

$$V_{e,\iota} = 62.91, \quad V_{e,c} = 59.23, \quad V_{e,o} = 70.02,$$

where the time indices of $y_{t\cdot\cdot}$ are omitted. The observation-panel design is rather inefficient in estimating the period-specific means as this design samples fewer units. Instead, the cluster-panel design is the most efficient one, closely followed by the independent design.

Figures 7-12 parallel Figure 1-6, respectively. Since the best design in terms of V is now the cluster-panel design, most of these plots show the preference toward this design. Figure 7 shows that when the variances of the contemporaneous means are taken into account, the simpler independent and cluster-panel designs are preferred for a greater fraction of parameter settings, and occupy a larger portion of the plot than in Figure 1. The point where the three designs are equivalent is $c_{12}^{I} = 20.6$, $c_{12}^{II} = 2.27$, closer to the origin than in Figure 1, in which only the variance of the difference was taken into account.

Figure 8 shows that the observation-panel design is only justified when both autocorrelations are higher than 0.6 (for the given values of population variances and costs). Recall that in Figure 2, the observation-panel design was preferred whenever $\rho^{II} > 0.34$, with little dependence on $\rho^{I}$.

Figure 9 shows how the PSU-level correlations and costs affect the choice of the design. The observation-panel design is less efficient than the cluster-panel design for all combinations of parameters in this plot. Hence, the choice of the design is between the independent and the cluster-panel designs. Naturally, if the data collection in the panel mode is expensive, the independent design is preferred to the cluster-panel design. Interestingly, the preference towards a particular design is not monotone in $\rho_{12}^{I}$. With values $\rho_{12}^{I} > 0.7$, the $V[d]$ component in (5.2) produces designs with so few clusters that $V[\bar{y}]$ suffers notably enough to hurt the whole objective function. At that value of panel autocorrelation, the maximum panel cost at which the cluster-panel design is still the most efficient one is $c_{12}^{I} = 24.4$, *i.e.*, the cluster-level cost in the second wave is 44% higher than in the first wave.

Figure 10 shows that the higher autocorrelation of the SSU measurements may justify modest extra cost associated with data collection. The highest cost for which the observation-panel design is still the most efficient one is $c_{12}^{II} = 2.75$ with $\rho^{II} = 0.78$; *i.e.*, the cost of the second interview can be 75% more than the cost of the first interview.

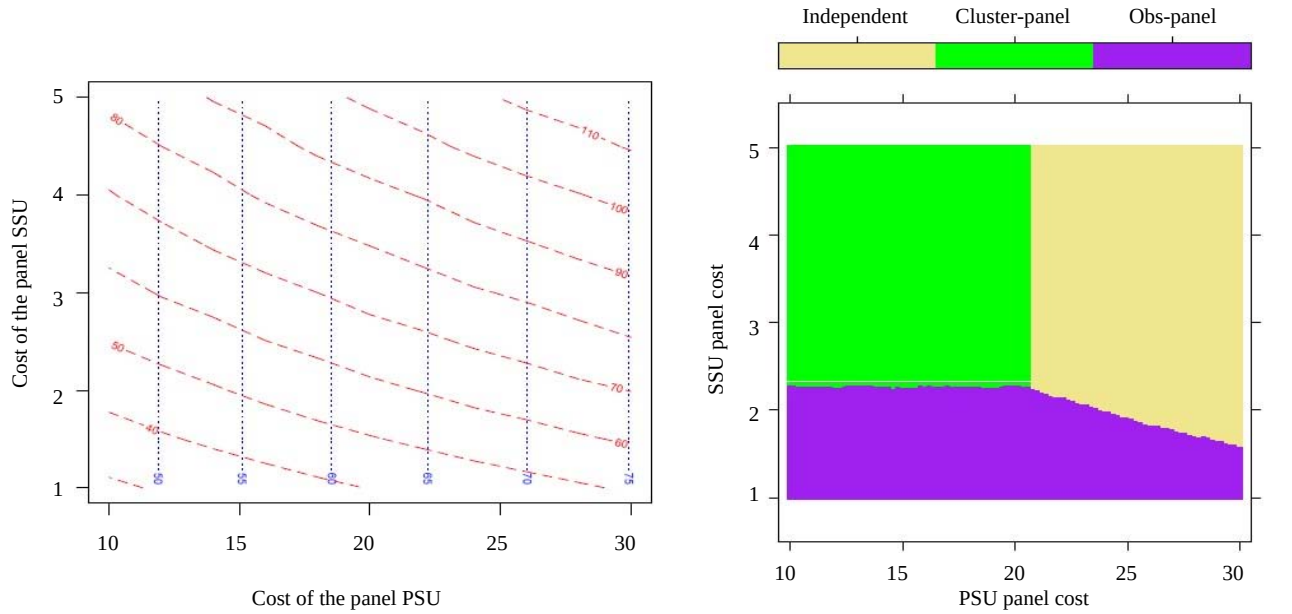


**Figure 7 Design variances as functions of the data collection costs $c_{12}^{I}, c_{12}^{II}$. Left: contour lines of $V_{e,c}$ (dotted) and $V_{e,o}$ (long dashed); $V_{e,\iota} = 62.91$; right: domains of optimality of the three designs**

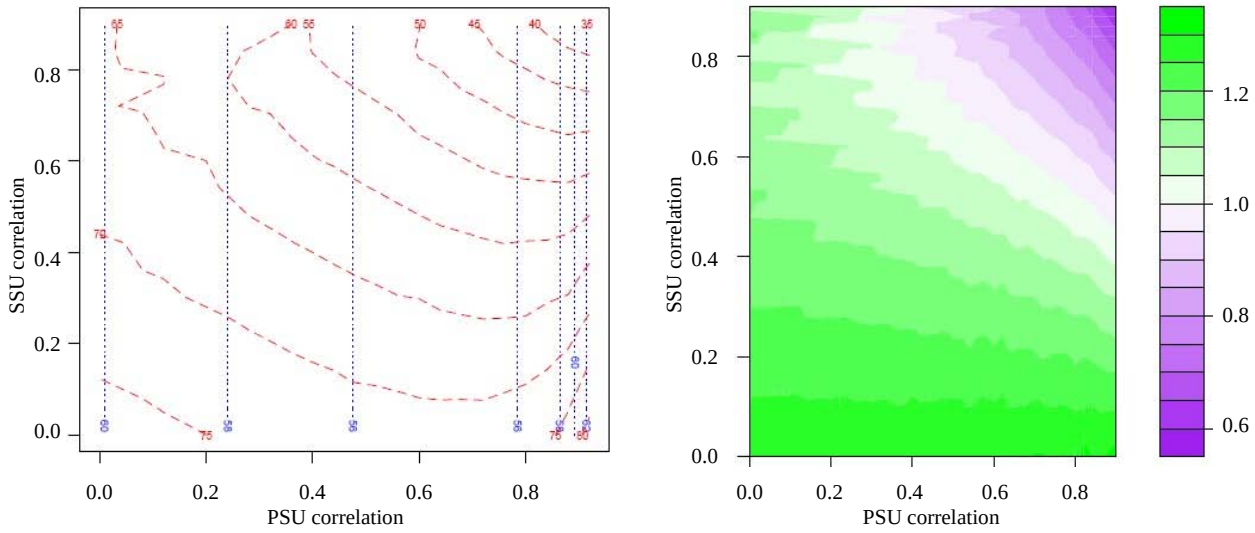


**Figure 8** Design variances as functions of the autocorrelations $\rho^{I}, \rho^{II}$. Left: contour lines of $V_{e,c}$ (dotted) and $V_{e,o}$ (long dashed); $V_{e,\iota} = 62.91$; right: ratio $V_{e,o}/V_{e,c}$

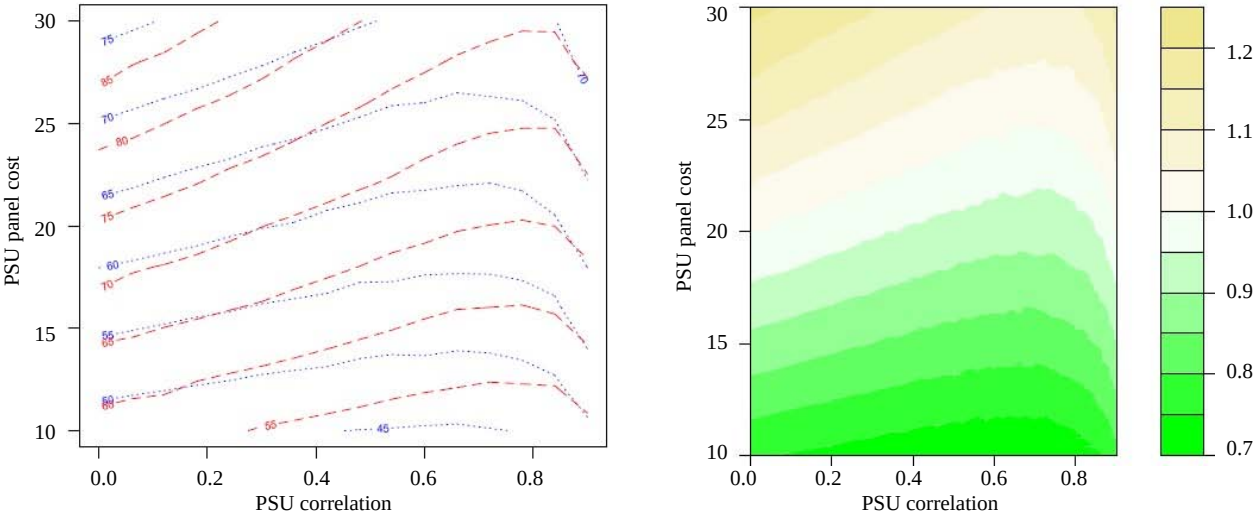


**Figure 9** Design variances as functions of the cluster-level autocorrelation $\rho^{I}$ and cost $c_{12}^{I}$. Left: contour lines of $V_{e,c}$ (dotted) and $V_{e,o}$ (long dashed); $V_{e,\iota} = 62.91$; right: ratio $V_{e,c}/V_{e,\iota}$

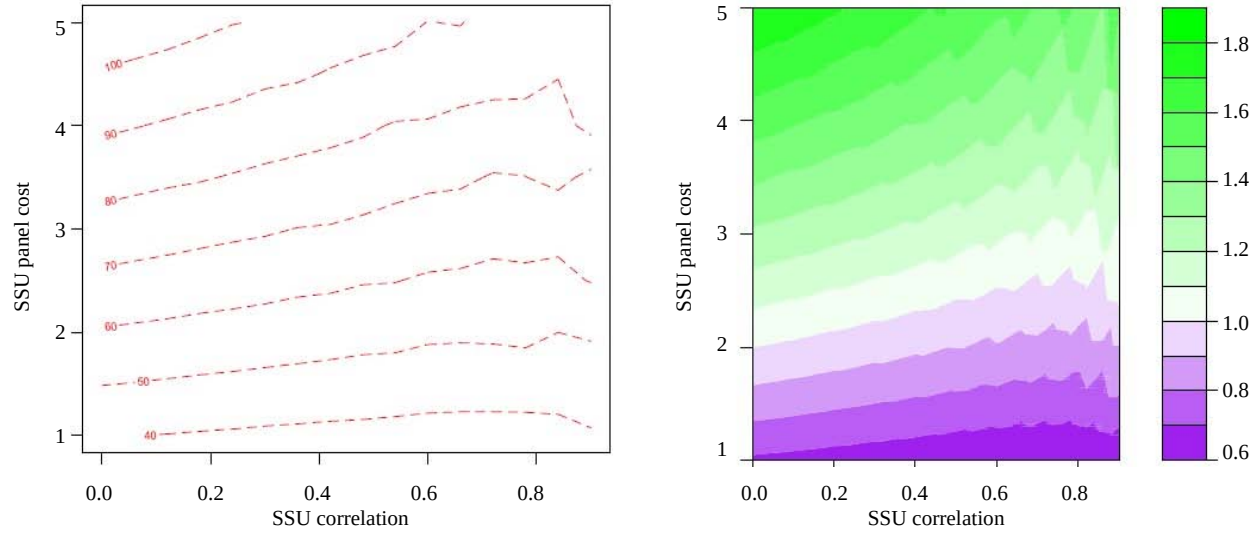


**Figure 10** Design variances as functions of the observation-level autocorrelation $\rho^{II}$ and cost $c_{12}^{II}$. Left: contour lines of $V_{e,o}$ (long dashed); $V_{e,\iota} = 62.91$; $V_{e,c} = 59.23$; right: ratio $V_{e,o}/V_{e,c}$

Figure 11 parallels Figure 5. The left panel shows that the observation-panel design is less efficient than the cluster-panel design. The right panel shows that if the cluster-level cost of the second wave exceeds the cluster-level cost of the first wave by more than 15 units, the independent design delivers better efficiency than the cluster-panel design.

Finally, Figure 12 shows the variances as functions of the total survey budget and the cost of the panel mode of data collection. There is very little dependence on $C_0$ in the plot, and the independent design is preferred if the panel mode is too expensive, namely, when the cluster-level cost in the second cost exceeds 107% of that in the first wave.

As it was conjectured in the beginning of this section, incorporation of the variances of the contemporaneous means into the design optimization objective function shifted the preferences of the survey designer towards simpler designs that can sample a greater number of the ultimate observation units. The observation-panel design now only makes sense when both the PSU and SSU autocorrelations are high, and the panel costs are reasonably low. Moreover, the cluster-panel design is generally justified only if there is an economy in cluster-level cost in the second wave of the survey.

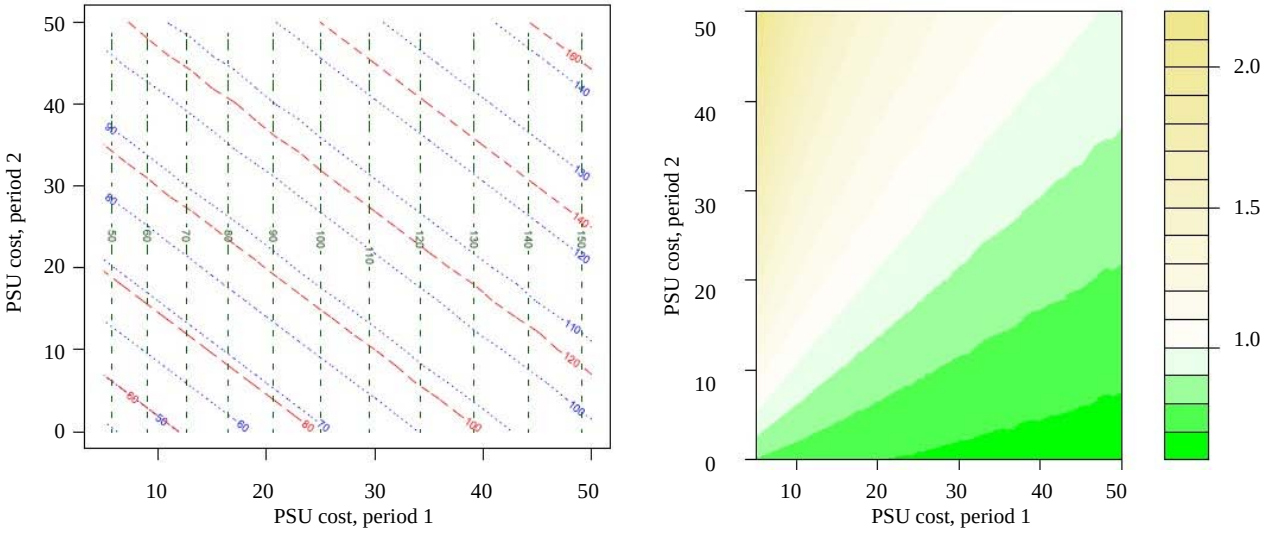


**Figure 11 Design variances as functions of the data collection costs $c_1^{\mathrm{I}}, c_{12}^{\mathrm{I}}$. Left: contour lines of $\mathrm{V}_{e,c}$ (dotted), $\mathrm{V}_{e,o}$ (long dashed) and $\mathrm{V}_{e,\iota}$ (dash-dotted); right: ratio $\mathrm{V}_{e,c}/\mathrm{V}_{e,\iota}$**

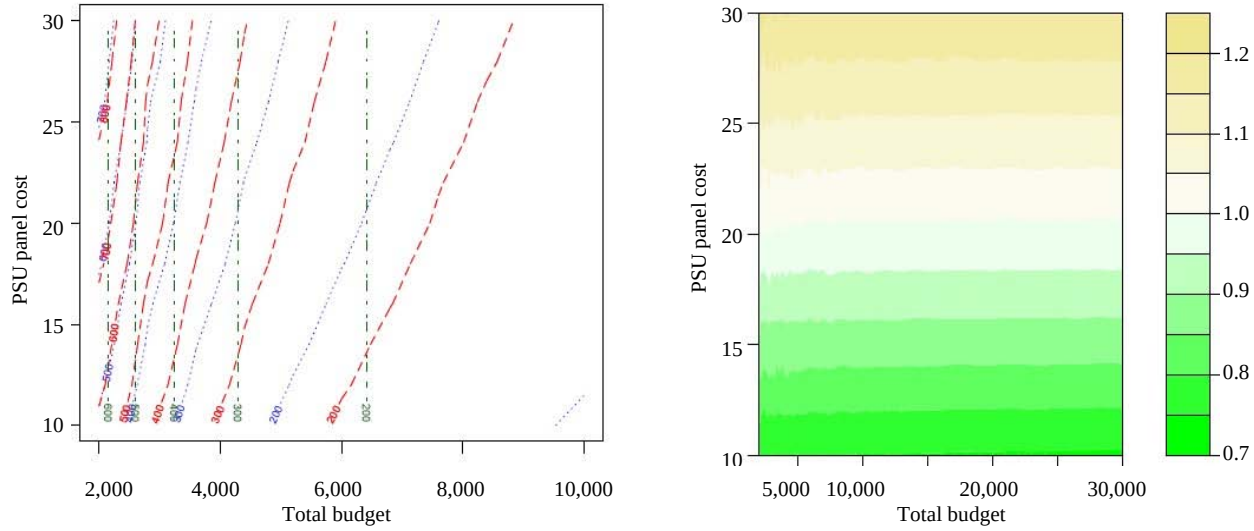


**Figure 12 Design variances as functions of the total budget $C_0$ and the PSU panel cost $c_{12}^{\mathrm{II}}$. Left: contour lines of $\mathrm{V}_{e,c}$ (dotted), $\mathrm{V}_{e,o}$ (long dashed) and $\mathrm{V}_{e,\iota}$ (dash-dotted); right: domains of optimality of the three designs**

## 6. Extensions to multiple waves

If the survey to be designed will have more than two waves of data collection, the survey designer may be able to extend the framework of the utility maximization problem (5.1), with the following considerations in mind.

1. A greater number of targets of inference. Possible variances that the survey designer may need to take into account can now include: contemporaneous variances $V[\bar{y}_1]$, $V[\bar{y}_2]$, …, $V[\bar{y}_T]$; consecutive differences $V[\bar{y}_2 - \bar{y}_1]$, …, $V[\bar{y}_T - \bar{y}_{T-1}]$ or composite/GLS estimators of the change between two adjacent periods of time; other contrasts $V[\sum_t c_t \bar{y}_t], \sum c_t = 0$; variance of the linear growth rates from regression of $\bar{y}_t$ on $t$, estimated by OLS or GLS; *etc.*
2. A possibility of discounting. In economics, it is customary to specify the budget constraints that look into the future in the form of $\sum_t x_t \delta^t$ where $x_t$ is the amount spent in time $t$, and $\delta < 1$ is the discount factor associated with interest rates. Discounting may also be relevant for the utility function, and design variances farther in the future may have lower weights in the optimization problem.
3. Unknown functional forms of the time-series processes associated with the variable of interest. The survey designer needs to have a good idea about the covariance structure of the time series of both individual observations and cluster means. It is likely that the results will be sensitive to the choice of the particular model. In the current analysis, the issue is ameliorated, as it suffices to have a single correlation parameter for each level. The survey designer may have to introduce more parameters into the model, and correspondingly study sensitivity of the design choice with respect to these parameters.

The complexity of the problem, as outlined above, can grow out of control very quickly. We thus abstain from a more detailed treatment of it in this paper.

## 7. Discussion

This paper has analyzed different options for implementation of repeated cluster surveys. We have provided analytical expression for design variances of the simple difference estimator for three popular designs (the independent, the cluster-panel and the observation-panel designs). We have also derived the optimal sample sizes for estimation of the difference between two waves of data collection.

The sample designer who knows that the characteristic of interest is going to have some degree of persistence over time will likely choose one of the panel designs, provided that the costs of re-visiting the clusters and/or observation units are not prohibitively high. Analytical comparison is possible between the independent and the observation-panel designs, and is given by Proposition 7. It is worth noting that the design variance of the difference is $O(C_0^{-1})$ for both the independent design and the observation-panel design, and is $O(C_0^{-1/2})$ for the cluster-panel design, where $C_0$ is the total budget of the survey. Hence the cluster-panel design is only viable for smaller surveys, while the large scale surveys will likely have either the independent or the observation-panel format.

The cost structure considered in Section 3 is rather simplistic. For instance, the second stage costs in the second time period may differ across individuals sampled from the new or from the reused clusters. Also, the costs may depend on the cluster size $M_i$, as it may take more time and resources to obtain maps and collect cluster level data for bigger clusters. Our original motivation was to consider situations in which the SSU panel cost is higher than twice the cost of individual interviews. However, as suggested by one of the referees, this cost may be lower if the follow-up interviews are performed in cheaper mode, such as a phone interview or a self-administered mail survey instead of a personal interview. If this is the case, the observation-panel design is apparently the most cost-efficient of the three designs.

The population structure is also an oversimplification. The clusters are assumed to be of balanced unchanging sizes. No units leave the population, and no new units appear. These assumptions are quite restrictive for many practical situations. If the population changes between two waves of data collection, the sample designer would want to include new clusters at the second wave, using the algorithms of Ernst (1999). The new clusters are placed into a separate stratum, and a clustered sample is taken from that stratum. In NHIS, this is implemented by "permit" frame. Also, the dynamic measurement effects such as conditioning and time in sample lead to rotation bias, so it might be beneficial to provide at least some rotation of the PSUs. For DHS studies, in particular, the first argument (coverage) is likely to be more important than the second one (time in sample) due to a substantial time between the waves of the survey (about 5 years). Arguably, both non-response and loss of coverage can be added to the current framework as sources of bias, leading to optimization of the mean squared total survey error rather than the design variance. Convincing models of such biases may be difficult to formulate, however.

Another issue that would arise with clusters of different sizes is that of the greater range of applicable designs. In this paper, we assumed SRSWOR at both stages. Other designs, such as sampling with probability proportional to size (PPS), can be used instead. For designs other than SRS, the Horvitz-Thompson estimator and its variance (Särndal, Swensson and Wretman 1992, Thompson 1997) would need to be used. The analytical derivations become unwieldy, although practical numerical demonstrations similar to our Sections 4 and 5 can still be implemented. If cluster sizes change over time, obtaining the optimal design becomes a moving target, and designs optimal for the "old" measures of size will lose their efficiency with the "new" measures of size.

In earlier drafts of this paper, we analyzed intermediate designs where a non-trivial fraction of the units are retained, and other units are sampled independently. The problem can then be viewed as variance minimization subject to inequality constraints on the degree of the overlap $0 \le \pi^{\mathrm{I}} \le 1$, $0 \le \pi^{\mathrm{II}} \le 1$. The general theory of non-linear constrained optimization ensures that as long as the variance of the population mean change $D$ is monotone in $\pi^{\mathrm{I}}$ and $\pi^{\mathrm{II}}$, the optimum will be achieved in one of the vertices of the parameter space. This justifies our interest in the three designs considered in the paper. They correspond to the vertices of the parameter space: $(0, 0)$, $(1, 0)$ and $(1, 1)$ for the independent, cluster-panel and observation-panel designs, respectively. The point $(0, 1)$ corresponds to an impossible design with complete overlap of the individual units with no overlap of the clusters. Cumbersome derivations show that it is possible to satisfy the first order conditions in some intermediate cases, too, but they correspond to local maxima of the variance. While these results may also be of interest (in the sense of providing an upper bound on the design variances), we did not consider them in the paper. In the more complicated cases of the multicriterial optimization of Section 5, monotonicity does not necessarily hold, and other designs beside the three extreme cases considered in the paper may lead to the optimal values of the objective function (5.2).

Conditions of equal variances (2.9) can be relaxed at the price of producing substantially more complicated expressions. If the sample sizes are fixed between the two occasions, then the following changes will be necessary in all relevant formulas. In the expressions that do not involve autocorrelations,

$$2S_b^2 \mapsto S_{1b}^2 + S_{2b}^2, \quad 2S_w^2 \mapsto \overline{S}_{1w}^2 + \overline{S}_{2w}^2, \qquad (7.1)$$

while in the expressions that do involve autocorrelations,

$$2(1-\rho^{\mathrm{I}})S_b^2 \mapsto S_{1b}^2 + S_{2b}^2 - 2\rho^{\mathrm{I}} S_{1b} S_{2b},$$
$$2S_w^2(1-\rho^{\mathrm{II}}) \mapsto \overline{S}_{1w}^2 + \overline{S}_{2w}^2 - 2\rho^{\mathrm{I}} \overline{S}_{1w} \overline{S}_{2w}. \qquad (7.2)$$

Qualitatively, the results will be the same.

The multicriterial framework of Section 5 allows for different importance weights to be given to different variances of interest. Relatively larger values of $\alpha_1, \alpha_2$ correspond to the greater importance of the contemporaneous means, while larger values of $\alpha_3$ correspond to the greater importance of the change estimate. The original problem of optimizing the design for $\mathrm{V}[d]$ can be considered within the context of (5.1) by setting $\alpha_1 = \alpha_2 = 0$, $\alpha_3 = 1$. This framework can also be expanded to include designs aimed at measuring several variables. An additional challenge of such a setup is that the autocorrelations may differ across different variables. Some individual characteristics are constant over time (race, gender); others change slowly (housing, expenditure, political preferences), yet others may change faster (income or behavior).

This paper dealt with three designs and a specific estimator of change: the difference in the two estimates of the mean in two periods of time. Other options for either designs or estimators are also available. For instance, in rotation designs, a fraction of the first wave units is retained, and some new units are recruited. For such designs, composite estimation (Hansen *et al.* 1953, Patterson 1950, Rao and Graham 1964, Wolter 2007) that weighs differently the contributions of the independent units (those retired from the sample after the first wave, and those newly recruited for the second wave) and the contributions of the panel units (used in both waves) would result in more efficient estimates. Generally, motivation for such designs comes from non-sampling considerations, such as decrease of the response burden and deterioration of the sample representativeness of population due to the population change. These considerations can be accounted for in either the cost model (*e.g.*, a greater number of callbacks required to convince a unit to respond), or the total survey error model (by introducing the non-response or undercoverage bias, and considering mean squared error rather than the design variance of an estimate).

## Acknowledgements

The authors are grateful to Chris Skinner and John Eltinge for helpful discussions, to William Kalsbeek for suggestions at the early stages of the paper, and to the associate editor and two referees for their comments. Nash Herndon and Oksana Loginova provided editorial improvements. Partial financial support was provided by U.S. Agency for International Development through the MEASURE Evaluation project of Carolina Population Center, University of North Carolina at Chapel Hill, under the terms of Cooperative Agreement GPO-A-00-03-00003-00. The authors are also grateful to the participants of the Joint Statistical Meetings

(2005) and the XXIII International Methodology Symposium of Statistics Canada (2007) for helpful comments.

## Appendix

Expectations, variances and covariances in the proofs below are with respect to the corresponding designs. The first stage of selection will be denoted with a superscript I. The second stage of selection will be denoted with a superscript II.

*Proof of Proposition* 2. Let us denote the sample of the PSUs by $\mathcal{S}^{\mathrm{I}}$, the sample of SSUs in the first period by $\mathcal{S}^{\mathrm{II}}_{i1}$, and the sample of SSUs in the second period by $\mathcal{S}^{\mathrm{II}}_{i2}$. Then

$$d = \bar{y}_{2\cdot\cdot} - \bar{y}_{1\cdot\cdot} = \frac{1}{mn}\sum_{i\in\mathcal{S}^{\mathrm{I}}}\left(\sum_{j\in\mathcal{S}^{\mathrm{II}}_{i2}} y_{2ij} - \sum_{j\in\mathcal{S}^{\mathrm{II}}_{i1}} y_{1ij}\right).$$

Denoting the expectations with respect to the first stage as $\mathrm{E}_{\mathrm{I}}$, and those with respect to the second stage as $\mathrm{E}_{\mathrm{II}}$, we have the design variance of $d$ equal to

$$\begin{aligned}
\mathrm{V}[d] &= \mathrm{E}_{\mathrm{I}}\mathrm{V}_{\mathrm{II}}[d \mid \mathcal{S}^{\mathrm{I}}] + \mathrm{V}_{\mathrm{I}}\mathrm{E}_{\mathrm{II}}[d \mid \mathcal{S}^{\mathrm{I}}] \\
&= \frac{1}{m^2n^2}\mathrm{E}_{\mathrm{I}}\left\{\sum_{i\in\mathcal{S}^{\mathrm{I}}}\mathrm{V}_{\mathrm{II}}\left[\sum_{j\in\mathcal{S}^{\mathrm{II}}_{i2}} y_{2ij} - \sum_{j\in\mathcal{S}^{\mathrm{II}}_{i1}} y_{1ij}\right]\right\} \\
&\quad + \frac{1}{m^2n^2}\mathrm{V}_{\mathrm{I}}\left\{\sum_{i\in\mathcal{S}^{\mathrm{I}}}\mathrm{E}_{\mathrm{II}}\left[\sum_{j\in\mathcal{S}^{\mathrm{II}}_{i2}} y_{2ij} - \sum_{j\in\mathcal{S}^{\mathrm{II}}_{i1}} y_{1ij}\right]\right\} \\
&= \frac{1}{m^2n^2}\mathrm{E}_{\mathrm{I}}\left\{\sum_{i\in\mathcal{S}^{\mathrm{I}}}\mathrm{V}_{\mathrm{II}}\left[\sum_{j\in\mathcal{S}^{\mathrm{II}}_{i2}} y_{2ij}\right] + \mathrm{V}_{\mathrm{II}}\left[\sum_{j\in\mathcal{S}^{\mathrm{II}}_{i1}} y_{1ij}\right]\right\} \\
&\quad + \frac{1}{m^2n^2}\mathrm{V}_{\mathrm{I}}\left[\sum_{i\in\mathcal{S}^{\mathrm{I}}} m\bar{Y}_{2i\cdot} - m\bar{Y}_{1i\cdot}\right] \\
&= \frac{1}{m^2n^2}\mathrm{E}_{\mathrm{I}}\left[\sum_{i\in\mathcal{S}^{\mathrm{I}}}\left(1-\frac{m}{M}\right)mS^2_{2wi} + \left(1-\frac{m}{M}\right)mS^2_{1wi}\right] \\
&\quad + \frac{1}{n^2}\left(1-\frac{n}{N}\right)n(S^2_{1b} + S^2_{2b} - 2\rho^{\mathrm{I}}S_{1b}S_{2b}) \\
&= \frac{1}{m^2n^2}nm\left(1-\frac{m}{M}\right)m(S^2_{2w} + S^2_{1w}) \\
&\quad + \frac{1}{n}\left(1-\frac{n}{N}\right)(S^2_{1b} + S^2_{2b} - 2\rho^{\mathrm{I}}S_{1b}S_{2b}) \\
&= \left(1-\frac{n}{N}\right)\frac{2S^2_b(1-\rho^{\mathrm{I}})}{n} + \left(1-\frac{m}{M}\right)\frac{2S^2_w}{mn},
\end{aligned}$$

where the last equality assumes symmetric conditions (2.9).

*Proof of Proposition* 3. Let us denote the sample of the PSUs by $\mathcal{S}^{\mathrm{I}}$, and the sample of SSUs, by $\mathcal{S}^{\mathrm{II}}_i$. Then

$$d = \bar{y}_{2\cdot\cdot} - \bar{y}_{1\cdot\cdot} = \frac{1}{mn}\sum_{i\in\mathcal{S}^{\mathrm{I}}}\sum_{j\in\mathcal{S}^{\mathrm{II}}_i}(y_{2ij} - y_{1ij}).$$

Denoting the expectations with respect to the first stage as $\mathrm{E}_{\mathrm{I}}$, and those with respect to the second stage as $\mathrm{E}_{\mathrm{II}}$, we have the design variance of $d$ equal to

$$\begin{aligned}
\mathrm{V}[d] &= \mathrm{E}_{\mathrm{I}}\mathrm{V}_{\mathrm{II}}[d \mid \mathcal{S}^{\mathrm{I}}] + \mathrm{V}_{\mathrm{I}}\mathrm{E}_{\mathrm{II}}[d \mid \mathcal{S}^{\mathrm{I}}] \\
&= \frac{1}{m^2n^2}\mathrm{E}_{\mathrm{I}}\left[\sum_{i\in\mathcal{S}^{\mathrm{I}}}\mathrm{V}_{\mathrm{II}}\sum_{j\in\mathcal{S}^{\mathrm{II}}_i}(y_{2ij} - y_{1ij})\right] \\
&\quad + \frac{1}{m^2n^2}\mathrm{V}_{\mathrm{I}}\left[\sum_{i\in\mathcal{S}^{\mathrm{I}}}\mathrm{E}_{\mathrm{II}}\sum_{j\in\mathcal{S}^{\mathrm{II}}_i}(y_{2ij} - y_{1ij})\right] \\
&= \frac{1}{m^2n^2}\mathrm{E}_{\mathrm{I}}m\left[\sum_{i\in\mathcal{S}^{\mathrm{I}}}\left(1-\frac{m}{M}\right)(S^2_{2wi}+S^2_{1wi}-2S_{2wi}S_{1wi}\rho^{\mathrm{II}})\right] \\
&\quad + \frac{1}{m^2n^2}\mathrm{V}_{\mathrm{I}}\left[\sum_{i\in\mathcal{S}^{\mathrm{I}}} m(\bar{Y}_{2ij} - \bar{Y}_{1ij})\right] \\
&= \frac{1}{mn^2}n\left(1-\frac{m}{M}\right)(\bar{S}^2_{1w} + \bar{S}^2_{2w} - 2\rho^{\mathrm{II}}\bar{S}_{1w}\bar{S}_{2w}) \\
&\quad + \frac{1}{n^2}n\left(1-\frac{n}{N}\right)(S^2_{1b} + S^2_{2b} - 2\rho^{\mathrm{I}}S_{1b}S_{2b}) \\
&= \left(1-\frac{n}{N}\right)\frac{S^2_{1b} + S^2_{2b} - 2\rho^{\mathrm{I}}S_{1b}S_{2b}}{n} \\
&\quad + \left(1-\frac{m}{M}\right)\frac{\bar{S}^2_{1w} + \bar{S}^2_{2w} - 2\rho^{\mathrm{II}}\bar{S}_{1w}\bar{S}_{2w}}{mn} \\
&= 2\left(1-\frac{n}{N}\right)\frac{S^2_b(1-\rho^{\mathrm{I}})}{n} \\
&\quad + 2\left(1-\frac{m}{M}\right)\frac{S^2_w(1-\rho^{\mathrm{II}})}{mn},
\end{aligned}$$

with the last equality holding under the symmetry conditions.

*Proof of Proposition* 4. The Lagrangian function of minimizing (2.11) subject to constraint (3.1) is

$$\begin{aligned}
L(n_1, m_1, n_2, m_2, \lambda) = &\left(1-\frac{n_1}{N}\right)\frac{S^2_b}{n_1} + \left(1-\frac{n_2}{N}\right)\frac{S^2_b}{n_2} \\
&+ \left(1-\frac{m_1}{M}\right)\frac{\bar{S}^2_w}{n_1m_1} + \left(1-\frac{m_2}{M}\right)\frac{\bar{S}^2_w}{n_2m_2} \\
&- \lambda(c^{\mathrm{I}}_1n_1 + c^{\mathrm{II}}_1n_1m_1 + c^{\mathrm{I}}_2n_2 + c^{\mathrm{II}}_2n_2m_2 - C_0).
\end{aligned}$$

Working through the first order conditions of this Lagrangian function leads to

$$-\lambda = \frac{m_1 S_b^2 + \left(1 - \frac{m_1}{M}\right)\bar{S}_w^2}{n_1^2 m_1 (c_1^{\mathrm{I}} + c_1^{\mathrm{II}} m_1)} = \frac{m_2 S_b^2 + \left(1 - \frac{m_2}{M}\right)\bar{S}_w^2}{n_2^2 m_2 (c_2^{\mathrm{I}} + c_1^{\mathrm{II}} m_2)}$$
$$= \frac{\bar{S}_w^2}{m_1^2 n_1^2 c_1^{\mathrm{II}}} = \frac{\bar{S}_w^2}{m_2^2 n_2^2 c_2^{\mathrm{II}}}.$$

Utilizing these conditions, we have

$$m^2 n^2 c^{\mathrm{II}} \left[ m S_b^2 + \left(1 - \frac{m}{M}\right)\bar{S}_w^2 \right] = n^2 m (c^{\mathrm{I}} + c^{\mathrm{II}} m)\bar{S}_w^2,$$

which can be written as

$$0 = (c^{\mathrm{I}} + c^{\mathrm{II}} m)\bar{S}_w^2 - m c^{\mathrm{II}} \left[ m S_b^2 + \left(1 - \frac{m}{M}\right)\bar{S}_w^2 \right]$$
$$= (c^{\mathrm{I}} + c^{\mathrm{II}} m) M \bar{S}_w^2 - m c^{\mathrm{II}} [M m S_b^2 + (M - m)\bar{S}_w^2]$$
$$= c^{\mathrm{I}} M \bar{S}_w^2 + m c^{\mathrm{II}} M \bar{S}_w^2 - m^2 c^{\mathrm{II}} M S_b^2 - m c^{\mathrm{II}} M \bar{S}_w^2 + m^2 c^{\mathrm{II}} \bar{S}_w^2$$
$$= c^{\mathrm{I}} M \bar{S}_w^2 + m^2 c^{\mathrm{II}} (\bar{S}_w^2 - M S_b^2).$$

Hence,

$$m = \sqrt{\frac{c^{\mathrm{I}}}{c^{\mathrm{II}}} \frac{\bar{S}_w^2}{S_b^2 - \bar{S}_w^2 / M}}.$$

From the survey budget (3.1), the number of clusters is found to be

$$n = \frac{C_0}{2(c^{\mathrm{I}} + m c^{\mathrm{II}})} = \frac{C_0}{2\{c^{\mathrm{I}} + [c^{\mathrm{I}} c^{\mathrm{II}} \bar{S}_w^2 / (S_b^2 - \bar{S}_w^2 / M)]^{1/2}\}}.$$

Plugging these expressions into (2.11) and using the equality relations (2.9), we obtain the variance of the estimator as

$$\mathrm{V}_{e,\iota}[d] = 2\left(1 - \frac{n}{N}\right)\frac{S_b^2}{n} + 2\left(1 - \frac{m}{M}\right)\frac{\bar{S}_w^2}{mn}$$
$$= 2\left[1 - \frac{C_0}{2(c^{\mathrm{I}} + m c^{\mathrm{II}})N}\right]\frac{2(c^{\mathrm{I}} + m c^{\mathrm{II}})S_b^2}{C_0}$$
$$+ 4\left(1 - \frac{m}{M}\right)\frac{\bar{S}_w^2 (c^{\mathrm{I}} + m c^{\mathrm{II}})}{m C_0}$$
$$= 2\left[\frac{2(c^{\mathrm{I}} + m c^{\mathrm{II}})}{C_0} - \frac{1}{N}\right]S_b^2$$
$$+ 4\left(\frac{1}{m} - \frac{1}{M}\right)\frac{\bar{S}_w^2 (c^{\mathrm{I}} + m c^{\mathrm{II}})}{C_0}$$
$$= \frac{4(c^{\mathrm{I}} + m c^{\mathrm{II}})}{C_0}\left[S_b^2 + \left(\frac{1}{m} - \frac{1}{M}\right)\bar{S}_w^2\right] - \frac{2}{N}S_b^2$$
$$= \frac{4\left[c^{\mathrm{I}} + \sqrt{c^{\mathrm{I}} c^{\mathrm{II}} \bar{S}_w^2 / (S_b^2 - \bar{S}_w^2 / M)}\right]}{C_0}$$
$$\times \left[S_b^2 + \left(\sqrt{\frac{c^{\mathrm{II}}}{c^{\mathrm{I}}}\frac{S_b^2 - \bar{S}_w^2 / M}{\bar{S}_w^2}} - \frac{1}{M}\right)\bar{S}_w^2\right] - \frac{2}{N}S_b^2.$$

*Proof of Proposition* 5. The Lagrangian function of minimizing (2.13) subject to constraint (3.4) is

$$L(n, m_1, m_2, \lambda) = 2\left(1 - \frac{n}{N}\right)\frac{(1 - \rho^{\mathrm{I}})S_b^2}{n}$$
$$+ \left(1 - \frac{m_1}{M}\right)\frac{\bar{S}_{1w}^2}{n m_1} + \left(1 - \frac{m_2}{M}\right)\frac{\bar{S}_{2w}^2}{n m_2}$$
$$- \lambda(c_{12}^{\mathrm{I}} n + c_1^{\mathrm{II}} n m_1 + c_2^{\mathrm{II}} n m_2 - C_0).$$

The first order conditions are:

$$\frac{\partial L}{\partial n} = -2\frac{(1 - \rho^{\mathrm{I}})S_b^2}{n^2} - \left(1 - \frac{m_1}{M}\right)\frac{\bar{S}_{1w}^2}{n^2 m_1}$$
$$- \left(1 - \frac{m_2}{M}\right)\frac{\bar{S}_{2w}^2}{n^2 m_2} - \lambda(c_{12}^{\mathrm{I}} + c_1^{\mathrm{II}} m_1 + c_2^{\mathrm{II}} m_2),$$
$$\frac{\partial L}{\partial m_1} = -\frac{\bar{S}_{1w}^2}{n m_1^2} - \lambda c_1^{\mathrm{II}} n,$$
$$\frac{\partial L}{\partial m_2} = -\frac{\bar{S}_{2w}^2}{n m_2^2} - \lambda c_2^{\mathrm{II}} n,$$
$$\frac{\partial L}{\partial \lambda} = c_{12}^{\mathrm{I}} n + c_1^{\mathrm{II}} n m_1 + c_2^{\mathrm{II}} n m_2 - C_0 = 0.$$

Expressing $-\lambda n$ from these conditions, one obtains:

$$-\lambda n = \frac{\bar{S}_{1w}^2}{m_1^2 n c_1^{\mathrm{II}}} = \frac{\bar{S}_{2w}^2}{m_2^2 n c_2^{\mathrm{II}}} = 2(1 - \rho^{\mathrm{I}})\frac{S_b^2}{C_0}$$
$$+ \left(1 - \frac{m_1}{M}\right)\frac{\bar{S}_{1w}^2}{m_1 C_0} + \left(1 - \frac{m_2}{M}\right)\frac{\bar{S}_{2w}^2}{m_2 C_0}.$$

Then

$$\frac{1}{m_2} = \frac{1}{m_1}\sqrt{\frac{c_2^{\mathrm{II}}}{c_1^{\mathrm{II}}}\frac{\bar{S}_{1w}^2}{\bar{S}_{2w}^2}} \equiv \frac{1}{\kappa m_1},$$
$$\frac{1}{m_1^2}\frac{(c_{12}^{\mathrm{I}} + c_1^{\mathrm{II}} m_1 + \kappa c_2^{\mathrm{II}} m_1)\bar{S}_{1w}^2}{c_1^{\mathrm{II}}} = 2(1 - \rho^{\mathrm{I}})S_b^2$$
$$+ \left(\frac{1}{m_1} - \frac{1}{M}\right)\bar{S}_{1w}^2 + \left(\frac{1}{\kappa m_1} - \frac{1}{\kappa M}\right)\bar{S}_{2w}^2,$$
$$0 = [2(1 - \rho^{\mathrm{I}})S_b^2 \kappa c_1^{\mathrm{II}} - \bar{S}_{1w}^2 \kappa c_1^{\mathrm{II}} / M - \bar{S}_{2w}^2 c_1^{\mathrm{II}} / M]m_1^2$$
$$+ [\bar{S}_{1w}^2 \kappa c_1^{\mathrm{II}} + \bar{S}_{2w}^2 c_1^{\mathrm{II}} - c_1^{\mathrm{II}} \bar{S}_{1w}^2 \kappa - \kappa^2 c_2^{\mathrm{II}} \bar{S}_{1w}^2]m_1 - c_{12}^{\mathrm{I}} \bar{S}_{1w}^2 \kappa,$$
$$D = [\bar{S}_{1w}^2 \kappa c_1^{\mathrm{II}} + \bar{S}_{2w}^2 c_1^{\mathrm{II}} - c_1^{\mathrm{II}} \bar{S}_{1w}^2 \kappa - \kappa^2 c_2^{\mathrm{II}} \bar{S}_{1w}^2]^2$$
$$+ 4[2(1 - \rho^{\mathrm{I}})S_b^2 \kappa - \bar{S}_{1w}^2 \kappa / M - \bar{S}_{2w}^2 / M]c_1^{\mathrm{II}} c_{12}^{\mathrm{I}} \bar{S}_{1w}^2 \geq 0,$$
$$m_1 = \frac{c_1^{\mathrm{II}} \bar{S}_{1w}^2 \kappa + \kappa^2 c_2^{\mathrm{II}} \bar{S}_{1w}^2 - \bar{S}_{1w}^2 \kappa c_1^{\mathrm{II}} - \bar{S}_{2w}^2 c_1^{\mathrm{II}} \pm \sqrt{D}}{4(1 - \rho^{\mathrm{I}})S_b^2 \kappa c_1^{\mathrm{II}} - 2\bar{S}_{1w}^2 \kappa c_1^{\mathrm{II}} / M - 2\bar{S}_{2w}^2 c_1^{\mathrm{II}} / M}.$$

The solution with $-\sqrt{D}$ leads to a negative value of $m_1$, and must be discarded.

The remaining design characteristics are

$$m_2 = \kappa m_1, \quad n = \frac{C_0}{c_{12}^{\mathrm{I}} + m_1 c_1^{\mathrm{II}} + m_2 c_1^{\mathrm{II}}}, \quad \kappa = \sqrt{\frac{c_1^{\mathrm{II}}}{c_2^{\mathrm{II}}}\frac{\bar{S}_{2w}^2}{\bar{S}_{1w}^2}}.$$

The variance of the difference estimator can be found using (2.15).

Under symmetric conditions, $\kappa = 1$, and

$$D = 4[2(1-\rho^{I})S_b^2 - 2\bar{S}_w^2/M]\, c^{II}c_{12}^{I}\bar{S}_w^2$$

is non-negative unless the expression in the square brackets is negative (which can only happen when $\rho^{I}$ is large and $M$ is small. In that case, a corner solution $m = M$ is realized). Furthermore,

$$m = m_1 = m_2 = \sqrt{\frac{\bar{S}_w^2 c_{12}^{I}}{2[(1-\rho^{I})S_b^2 - \bar{S}_w^2/M]c^{II}}},$$

$$n = \frac{C_0}{c_{12}^{I} + 2mc^{II}} = \frac{C_0}{c_{12}^{I} + \sqrt{\dfrac{2\bar{S}_w^2 c_{12}^{I} c^{II}}{(1-\rho^{I})S_b^2 - \bar{S}_w^2/M}}},$$

$$\begin{aligned}
&\mathrm{V}_{e,o}[d] \\
&= 2\left(1-\frac{n}{N}\right)\frac{(1-\rho^{I})S_b^2}{n} + 2\left(1-\frac{m}{M}\right)\frac{(1-\rho^{II})\bar{S}_w^2}{nm} \\
&= \frac{2}{n}\left[(1-\rho^{I})S_b^2 + 2\left(1-\frac{m}{M}\right)\frac{(1-\rho^{II})\bar{S}_w^2}{m}\right] - \frac{2(1-\rho^{I})S_b^2}{N} \\
&= \frac{2}{C_0}(c_{12}^{I} + 2mc^{II}) \\
&\times\left[(1-\rho^{I})S_b^2 + 2\left(\frac{1}{m}-\frac{1}{M}\right)(1-\rho^{II})\bar{S}_w^2\right] - \frac{2(1-\rho^{I})S_b^2}{N} \\
&= \frac{2}{C_0}(c_{12}^{I} + 2mc^{II}) \\
&\times\left[(1-\rho^{I})S_b^2 + \frac{2}{m}(1-\rho^{II})\bar{S}_w^2 - \frac{2}{M}(1-\rho^{II})\bar{S}_w^2\right] - \frac{2(1-\rho^{I})S_b^2}{N} \\
&= \frac{2}{C_0}\left\{c_{12}^{I}(1-\rho^{I})S_b^2 + 2(1-\rho^{II})\bar{S}_w^2\left[2c^{II} - \frac{c_{12}^{I}}{M}\right]\right. \\
&\quad \left. + \frac{2}{m}c_{12}^{I}(1-\rho^{II})\bar{S}_w^2 + 2mc^{II}\left[(1-\rho^{I})S_b^2 - \frac{2}{M}(1-\rho^{II})\bar{S}_w^2\right]\right\} \\
&\quad - \frac{2(1-\rho^{I})S_b^2}{N} \\
&= \frac{2}{C_0}\left\{c_{12}^{I}(1-\rho^{I})S_b^2 + 2(1-\rho^{II})\bar{S}_w^2\left[2c^{II} - \frac{c_{12}^{I}}{M}\right]\right. \\
&\quad + 2(1-\rho^{II})\sqrt{2[(1-\rho^{I})S_b^2 - \bar{S}_w^2/M]\bar{S}_w^2 c^{II}c_{12}^{I}} \\
&\quad \left. + \sqrt{\frac{2\bar{S}_w^2 c_{12}^{I}c^{II}}{(1-\rho^{I})S_b^2 - \bar{S}_w^2/M}}\left[(1-\rho^{I})S_b^2 - \frac{2}{M}(1-\rho^{II})\bar{S}_w^2\right]\right\} \\
&\quad - \frac{2(1-\rho^{I})S_b^2}{N}.
\end{aligned}$$

*Proof of Proposition* 6. The Lagrangian function of minimizing (2.15) subject to constraint (3.6) is

$$\begin{aligned}
L(n, m, \lambda) &= 2\left(1-\frac{n}{N}\right)\frac{(1-\rho^{I})S_b^2}{n} + 2\left(1-\frac{m}{M}\right)\frac{(1-\rho^{II})\bar{S}_w^2}{nm} \\
&\quad - \lambda(c_{12}^{I}n + c_{12}^{II}nm - C_0).
\end{aligned}$$

The first order conditions are:

$$\begin{aligned}
0 &= \frac{\partial L}{\partial n} = -2\frac{(1-\rho^{I})S_b^2}{n^2} \\
&\quad - 2\left(1-\frac{m}{M}\right)\frac{(1-\rho^{II})\bar{S}_w^2}{n^2 m} - \lambda(c_{12}^{I} + c_{12}^{II}m), \\
0 &= \frac{\partial L}{\partial m} = -2\frac{(1-\rho^{II})\bar{S}_w^2}{nm^2} - \lambda c_{12}^{II}n, \\
0 &= \frac{\partial L}{\partial \lambda} = c_{12}^{I}n + c_{12}^{II}nm - C_0.
\end{aligned}$$

Expressing $-\lambda n^2$ from these conditions, one obtains:

$$\begin{aligned}
-\lambda n^2/2 &= \frac{(1-\rho^{I})S_b^2}{c_{12}^{I} + c_{12}^{II}m} + \left(1-\frac{m}{M}\right)\frac{(1-\rho^{II})\bar{S}_w^2}{m(c_{12}^{I} + c_{12}^{II}m)} \\
&= \frac{(1-\rho^{II})\bar{S}_w^2}{m^2 c_{12}^{II}}.
\end{aligned}$$

Hence,

$$\begin{aligned}
&(1-\rho^{I})S_b^2 Mm^2 c_{12}^{II} + (M-m)(1-\rho^{II})\bar{S}_w^2 mc_{12}^{II} \\
&\qquad - (1-\rho^{II})M\bar{S}_w^2(c_{12}^{I} + c_{12}^{II}m) = 0, \\
&[(1-\rho^{I})S_b^2 Mc_{12}^{II} - (1-\rho^{II})\bar{S}_w^2 c_{12}^{II}]m^2 \\
&\qquad - [(1-\rho^{II})M\bar{S}_w^2 c_{12}^{I}] = 0, \\
m &= \sqrt{\frac{(1-\rho^{II})M\bar{S}_w^2 c_{12}^{I}}{[(1-\rho^{I})S_b^2 M - (1-\rho^{II})\bar{S}_w^2]c_{12}^{II}}} \\
&= \sqrt{\frac{c_{12}^{I}}{c_{12}^{II}}\frac{(1-\rho^{II})\bar{S}_w^2}{(1-\rho^{I})S_b^2 - (1-\rho^{II})\bar{S}_w^2/M}}.
\end{aligned}$$

From the survey budget (3.6),

$$n = \frac{C_0}{c_{12}^{I} + c_{12}^{II}m} = \frac{C_0}{c_{12}^{I} + \sqrt{\dfrac{(1-\rho^{II})\bar{S}_w^2 c_{12}^{I}c_{12}^{II}}{(1-\rho^{I})S_b^2 - (1-\rho^{II})\bar{S}_w^2/M}}}.$$

Finally, the variance of the difference estimator is

$$
\begin{aligned}
&\mathrm{V}_{e,o}[d] \\
&= \frac{2}{C_0}\left(c_{12}^{\mathrm{I}} + \sqrt{\frac{(1-\rho^{\mathrm{II}})\bar{S}_w^2 c_{12}^{\mathrm{I}} c_{12}^{\mathrm{II}}}{(1-\rho^{\mathrm{I}})S_b^2 - (1-\rho^{\mathrm{II}})\bar{S}_w^2 / M}}\right) \\
&\quad \times \left[(1-\rho^{\mathrm{I}})S_b^2 \right. \\
&\qquad \left. + \left(\sqrt{\frac{c_{12}^{\mathrm{II}}}{c_{12}^{\mathrm{I}}}\frac{(1-\rho^{\mathrm{I}})S_b^2 - (1-\rho^{\mathrm{II}})\bar{S}_w^2 / M}{(1-\rho^{\mathrm{II}})\bar{S}_w^2}} - \frac{1}{M}\right)(1-\rho^{\mathrm{II}})\bar{S}_w^2\right] \\
&\quad - \frac{2(1-\rho^{\mathrm{I}})S_b^2}{N} \\
&= \frac{2}{C_0}\left\{(1-\rho^{\mathrm{I}})S_b^2 c_{12}^{\mathrm{I}}\right. \\
&\qquad + (1-\rho^{\mathrm{II}})\bar{S}_w^2\sqrt{c_{12}^{\mathrm{I}}c_{12}^{\mathrm{II}}\frac{(1-\rho^{\mathrm{I}})S_b^2 - (1-\rho^{\mathrm{II}})\bar{S}_w^2 / M}{(1-\rho^{\mathrm{II}})\bar{S}_w^2}} \\
&\qquad + \left[(1-\rho^{\mathrm{I}})S_b^2 - \frac{1}{M}(1-\rho^{\mathrm{II}})\bar{S}_w^2\right] \\
&\qquad \times \sqrt{\frac{(1-\rho^{\mathrm{II}})\bar{S}_w^2 c_{12}^{\mathrm{I}} c_{12}^{\mathrm{II}}}{(1-\rho^{\mathrm{I}})S_b^2 - (1-\rho^{\mathrm{II}})\bar{S}_w^2 / M}} \\
&\qquad \left. + (1-\rho^{\mathrm{II}})\bar{S}_w^2\left(c_{12}^{\mathrm{II}} - \frac{c_{12}^{\mathrm{I}}}{M}\right)\right\} \\
&\quad - \frac{2(1-\rho^{\mathrm{I}})S_b^2}{N}.
\end{aligned}
$$

*Proof of Proposition* 7. Ignoring the finite population correcting terms of the order $O(N^{-1})$ and $O(M^{-1})$, equation (3.3) can be written as:

$$
\begin{aligned}
\mathrm{V}_{e,i}[d] &\approx \frac{4\left(c^{\mathrm{I}} + \sqrt{c^{\mathrm{I}}c^{\mathrm{II}}\bar{S}_w^2/S_b^2}\right)}{C_0}\left[S_b^2 + \left(\sqrt{\frac{c^{\mathrm{II}}}{c^{\mathrm{I}}}}\bar{S}_w^2 S_b^2\right)\right] \\
&= \frac{4}{C_0}\left(c^{\mathrm{I}}S_b^2 + c^{\mathrm{II}}\bar{S}_w^2 + 2\sqrt{c^{\mathrm{I}}c^{\mathrm{II}}S_b^2\bar{S}_w^2}\right) \\
&= \frac{4}{C_0}\left(\sqrt{c^{\mathrm{I}}S_b^2} + \sqrt{c^{\mathrm{II}}\bar{S}_w^2}\right)^2.
\end{aligned}
$$

Likewise, equation (3.8) can be written as

$$
\begin{aligned}
\mathrm{V}_{e,o}[d] &\approx \frac{2}{C_0}\left[(1-\rho^{\mathrm{I}})S_b^2 c_{12}^{\mathrm{I}}\right. \\
&\qquad \left. + 2\sqrt{c_{12}^{\mathrm{I}}c_{12}^{\mathrm{II}}(1-\rho^{\mathrm{I}})S_b^2(1-\rho^{\mathrm{II}})\bar{S}_w^2} + (1-\rho^{\mathrm{II}})\bar{S}_w^2 c_{12}^{\mathrm{II}}\right] \\
&= \frac{2}{C_0}\left[\sqrt{(1-\rho^{\mathrm{I}})S_b^2 c_{12}^{\mathrm{I}}} + \sqrt{(1-\rho^{\mathrm{II}})\bar{S}_w^2 c_{12}^{\mathrm{II}}}\right]^2.
\end{aligned}
$$

The statement of Propostion 7 follows immediately from these two expressions.

## References


Binder, D.A., and Hidiroglou, M.A. (1988). Sampling in time. In *Handbook of Statistics*, (Eds., P.R. Krishnaiah and C.R. Rao), North Holland, Amsterdam, 6, 187-211.

Cochran, W.G. (1977). *Sampling Techniques*, 3rd Ed., New York: John Wiley & Sons, Inc.

Eckler, A.R. (1955). Rotation sampling. *Annals of Mathematical Statistics*, 26(4), 664-685.

Ernst, L.R. (1999). The maximization and minimization of sample overlap problems: A half century of results. Technical report, U.S. Bureau of Labor Statistics.

Fuller, W.A. (1999). Environmental surveys over time. *Journal of Agricultural, Biological and Environmental Statistics*, 4(4), 331-345.

Groves, R.M. (1989). *Survey Errors and Survey Costs*. New York: John Wiley & Sons, Inc.

Hansen, M., Hurwitz, W.N. and Madow, W.G. (1953). *Sample Survey Methods and Theory*. New York: John Wiley & Sons, Inc.

Kish, L. (1995). *Survey Sampling*, 3rd Ed., New York: John Wiley & Sons, Inc.

Lehtonen, R., and Pahkinen, E. (2004). *Practical Methods for Design and Analysis of Complex Surveys*, *Statistics in Practice*, 2nd Ed., New York: John Wiley & Sons, Inc.

Mas-Colell, A., Whinston, M.D. and Green, J.R. (1995). *Microeconomic Theory*, Oxford University Press, Oxford, UK.

McDonald, T.L. (2003). Review of environmental monitoring methods: Survey designs. *Environmental Monitoring and Assessment*, 85, 277-292.

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

Patterson, H.D. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society*, Series B, 12(2), 241-255.

Rao, J.N.K., and Graham, J.E. (1964). Rotation designs for sampling on repeated occasions. *Journal of the American Statistical Association*, 59(306), 492-509.

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

Scott, C.T. (1998). Sampling methods for estimating change in forest resources. *Ecological Applications*, 8(2), 228-233.

Thompson, M.E. (1997). Theory of Sample Surveys. *Monographs on Statistics and Applied Probability*, New York: Chapman & Hall/CRC, 74.

Thompson, S.K. (1992). *Sampling*, New York: John Wiley & Sons, Inc.

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

Survey Methodology, June 2011
Vol. 37, No. 1, pp. 95-102
Statistics Canada, Catalogue No. 12-001-X



# On the efficiency of randomized probability proportional to size sampling

**Paul Knottnerus** [1]

Abstract

This paper examines the efficiency of the Horvitz-Thompson estimator from a systematic probability proportional to size (PPS) sample drawn from a randomly ordered list. In particular, the efficiency is compared with that of an ordinary ratio estimator. The theoretical results are confirmed empirically with of a simulation study using Dutch data from the Producer Price Index.

Key Words: Horvitz-Thompson estimator; Producer Price Index; Ratio estimator; Sampling autocorrelation coefficient.

## 1. Introduction

When the study variable $y$ in a population of $N$ units is more or less proportional to a size variable $x$, one may use the ratio estimator from a simple random sample of size $n$ without replacement (SRS). An alternative estimator in such a situation is the Horvitz-Thompson (HT) estimator in combination with a systematic probability proportional to size sample from a randomly ordered list, henceforth called a randomized PPS sample.

In recent years several authors investigated variance estimation procedures for the HT estimator from a randomized PPS sample. See, among others, Brewer and Donadio (2003), Cumberland and Royall (1981), Deville (1999), Knottnerus (2003), Kott (1988 and 2005), Rosén (1997) and Stehman and Overton (1994). For a comparison between the efficiencies of the ratio estimator and the randomized PPS estimator, the reader is referred to Foreman and Brewer (1971), Cochran (1977) and the references given therein. A drawback of these comparisons is that finite populations corrections are ignored. Hartley and Rao (1962) take the finite population correction into account but without an explicit formula for the efficiency. Elaborating on the results of Gabler (1984), Qualité (2008) shows that the related HT estimator from a rejective Poisson sample of size $n$ is more efficient than the Hansen-Hurwitz estimator for a sampling scheme with replacement. No formula for the increased efficiency is given, however.

The main aim of this paper is to derive formulas for the efficiency of the randomized PPS estimator relative to the ratio estimator. To this end, we present a simple formula for the change in the sample size required to maintain the same variance when a randomized PPS estimator is replaced by a ratio estimator. From the design based point of view these formulas are valid when $n = o(N)$ as $N \to \infty$. This condition suggests that the finite population correction can be neglected for this kind of sampling design. Surprisingly, as we will see in an example in section 4, the randomized PPS sampling can reduce variance by more than 30% compared to PPS sampling *with* replacement even when the sampling fraction $n/N$ is much smaller than 30%; see also Kott (2005, page 436). Furthermore, the formulas remain appropriate from a model assisted point of view when $n$ and $N$ are of the same order, provided that $N$ is large and that the hypothetical model for the observations $Y_i$ ($i = 1, ..., N$) satisfies mild conditions.

The outline of the paper is as follows. Section 2 describes an alternative expression for the variance of the HT estimator based on the sampling autocorrelation coefficient. The corresponding variance estimator for randomized PPS sampling is shown to be nonnegative with probability 1. Section 3 presents the formulas for the efficiency of the randomized PPS estimator relative to the ratio estimator for various data patterns often met in practice. Section 4 features an example with data on the Producer Price Index in The Netherlands illustrating the substantial efficiency gains obtainable in practice. A counterexample shows that randomized PPS sampling is not *always* advantageous. The paper concludes with a summary.

## 2. An alternative variance expression for randomized PPS sampling

Consider a population $U = \{1, ..., N\}$, and let $s$ be a sample of fixed size $n$ drawn from $U$ without replacement according to a given sampling design with first order inclusion probabilities $\pi_i$ and second order inclusion probabilities $\pi_{ij}$ ($i, j = 1, ..., N$). The HT estimator of the population total, $Y = \Sigma_{i\in U} Y_i$, is defined by $\hat{Y}_{\mathrm{HT}} = \Sigma_{i\in s} Y_i/\pi_i$. Suppose there is a measure of relative size $X_i$ (*i.e.*, $X = \Sigma_{i\in U} X_i = 1$) such that all $X_i \le 1/n$. In fact, it is assumed here that units with $X_i > 1/n$ are put together in a separate certainty-stratum. When the $\pi_i$ are proportional to these size measures, $\pi_i = nX_i$. Defining $Z_i = Y_i/X_i$, we can write $Y$ as a weighted mean of the $Z_i$, that is, $Y = \mu_z = \Sigma_{i\in U} X_i Z_i$. Likewise, we can write the HT estimator of $Y$ in randomized PPS sampling as $\hat{Y}_{\mathrm{HT}} = \hat{Y}_{\mathrm{PPS}} = \bar{z}_s$, where $\bar{z}_s$ is sample mean of the $Z_i$.

1. Paul Knottnerus, Statistics Netherlands, PO Box 24500, 2490 HA The Hague, The Netherlands. E-mail: pkts@cbs.nl.

The variance of the randomized PPS estimator $\hat{Y}_{\text{PPS}}$ is

$$\text{var}(\hat{Y}_{\text{PPS}}) = \frac{1}{n^2}\sum_{i\in U}\sum_{j\in U}(\pi_{ij} - \pi_i\pi_j)Z_iZ_j \quad (1)$$

$$= -\frac{1}{2n^2}\sum_{i\in U}\sum_{j\in U}(\pi_{ij} - \pi_i\pi_j)(Z_i - Z_j)^2 \quad (2)$$

with $\pi_{ii} = \pi_i$. The former is attributed to Horvitz and Thompson (1952) and the latter is due to Sen (1953) and Yates and Grundy (1953). The following alternative expression for the variance is more convenient for our purposes:

$$\text{var}(\hat{Y}_{\text{PPS}}) = \text{var}(\bar{z}_s) = \{1 + (n-1)\rho_z\}\frac{\sigma_z^2}{n}, \quad (3)$$

where $\sigma_z^2 = \sum_{i\in U} X_i(Z_i - \mu_z)^2$, and

$$\rho_z = \sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}}\frac{\pi_{ij}}{n(n-1)}\left(\frac{Z_i - \mu_z}{\sigma_z}\right)\left(\frac{Z_j - \mu_z}{\sigma_z}\right). \quad (4)$$

For a proof of (3), see Knottnerus (2003, page 103). Note that $\sigma_z^2/n$ would have been the variance if the sample had been drawn with replacement with drawing probabilities $X_i$.

The sampling autocorrelation coefficient $\rho_z$ in (4) is a generalization of the more familiar intraclass correlation coefficient $\rho$ in systematic sampling with equal probabilities; see, for instance, Cochran (1977, pages 209 and 240) and Särndal, Swensson and Wretman (1992, page 79). Note that $\rho_z$ is a fixed population parameter. The phrase *sampling autocorrelation* is used because $\rho_z$ refers to the autocorrelation between two randomly chosen observations, say $z_{s1}$ and $z_{s2}$, from $s$. Consequently, the value of $\rho_z$ depends on the sampling design. In particular, when sampling with replacement, $\rho_z = 0$, while under SRS sampling, $\rho_z = -1/(N-1)$.

Although exact expressions for the $\pi_{ij}$ under randomized PPS sampling are available, they can be cumbersome when $N$ is large. For an exact expression, see Connor (1966) and for a modification Hidiroglou and Gray (1980). Here we use an approximation proposed by Knottnerus (2003, page 197):

$$\pi_{ijK} = n(n-1)\frac{X_iX_j(1 - X_i - X_j)}{\gamma(1-2X_i)(1-2X_j)} \quad (5)$$

$$\gamma = \frac{1}{2} + \frac{1}{2}\sum_{i\in U}\frac{X_i}{1-2X_i}.$$

These $\pi_{ijK}$ have been shown to satisfy the second-order restrictions for the $\pi_{ij}$:

$$\sum_{i,\, j\in U(j\neq i)}\pi_{ij} = n(n-1),$$

and

$$\sum_{j\in U(j\neq i)}\pi_{ij} = (n-1)\pi_i.$$

Furthermore, (5) is correct for SRS sampling for any $n \leq N$, while $\pi_{ijK}$ coincide with the $\pi_{ijBD}$ from the special designs proposed by Brewer (1963a) and Durbin (1967) for PPS samples with $n = 2$. Moreover, the $\pi_{ijK}$ in (5) can be written in factorized form as proposed by Brewer and Donadio (2003). That is,

$$\pi_{ijK} = \pi_i\pi_j(c_i + c_j)/2, \quad (6)$$

and

$$c_i = (n-1)/n\gamma(1 - 2X_i).$$

An implication of approximation (5) is that $\pi_{ijK}/n(n-1)$ does not depend on $n$. Hence, the corresponding approximation of $\rho_z$ does not depend on $n$ (recall we have assumed that every $X_i < 1/n$).

This nondependence on $n$ would also result had we used the approximation proposed by Hartley and Rao (1962) for randomized PPS sampling:

$$\pi_{ijHR} = n(n-1)X_iX_j$$
$$\{1 + X_i + X_j - \mu_x + 2(X_i^2 + X_j^2 + X_iX_j)$$
$$- 3\mu_x(X_i + X_j - \mu_x - 2\Sigma_{i\in U}X_i^3)\}, \quad (7)$$

where $\mu_x = \Sigma_{i\in U}X_i^2$ (recall $\mu_z = \Sigma_{i\in U}X_iZ_i$). Obviously, $\pi_{ijHR}/n(n-1)$ does not depend on $n$. At the time Hartley and Rao assumed that $n = O(1)$ as $N \to \infty$. In addition, referring to a private conversation with J.N.K. Rao, Thompson and Wu (2008) state that approximation (7) is valid when $n = o(N)$ as $N \to \infty$. For an example that (5) and (7) can not be used for *any* $n$ and $N$, see Appendix A.

Since both (5) and (7) lead to approximations for $\rho_z$ in randomized PPS sampling that are $\rho_z\{1+o(1)\}$ as $N \to \infty$ with $n = o(N)$, (5) can be used for calculating $\rho_z$ in practice when $n << N$ and $N$ is large. For ease of the exposition, it is assumed here that there is a positive constant $c$ such that $\rho_z < -c/N$. See also Kott (2005, page 436) who discusses estimating the variance under PPS sampling when $n = O(N^{2/3})$.

Suppose $\gamma = 1 + \mu_x + O(1/N^2)$ and $\mu_x = O(1/N)$ (which follow from the conditions of Theorem 1 below). It is not hard to see that, after dropping $O(1/nN)$ terms, $c_i$ in (6) is identical with $c_{iHR} = (n-1)/\{n(1 + \mu_x - 2X_i)\}$. The latter expression is equation (11) of Brewer and Donadio, which is based on $\pi_{ijHR}$ in (7).

The approach proposed here is somewhat different from Knottnerus (2003). First, rewrite (5) as

$$\pi_{ijK} = n(n-1)\frac{X_iX_j}{\gamma}\left(\frac{1/2}{1-2X_i} + \frac{1/2}{1-2X_j}\right). \quad (8)$$

Substituting (8) into (4), we obtain a new, simple approximation for $\rho_z$:

$$\rho_z = \sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}} \frac{X_i X_j}{\gamma}\left(\frac{1/2}{1-2X_i}+\frac{1/2}{1-2X_j}\right)\left(\frac{Z_i-Y}{\sigma_z}\right)\left(\frac{Z_j-Y}{\sigma_z}\right)$$

$$= \sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}} \frac{X_i X_j}{\gamma}\left(\frac{1}{1-2X_i}\right)\left(\frac{Z_i-Y}{\sigma_z}\right)\left(\frac{Z_j-Y}{\sigma_z}\right)$$

$$= 0 - \sum_{i\in U}\frac{X_i^2}{\gamma(1-2X_i)}\left(\frac{Z_i-Y}{\sigma_z}\right)^2. \qquad (9)$$

In the second line, we used the equality $\sum_{i,j} m_{ij} v_i = \sum_{i,j} m_{ij} v_j$ when $m_{ij} = m_{ji}$. In the last line, we used $\sum_{j\in U} X_j (Z_j - Y) = 0$.

Next, let $\bar{X}$ denote the population mean of $X_1, ..., X_N$ and define $\sigma_x^2$ and $V_x^2$ by

$$\sigma_x^2 = \sum_{i\in U} X_i (X_i - \mu_x)^2,$$

and

$$V_x^2 = \sum_{i\in U} (X_i - \bar{X})^2 / N,$$

respectively. In the following theorem (9) is further simplified.

*Theorem* 1. Suppose that $(Z_i - Y)/\sigma_z = O(1)$ as $N \to \infty$ and that there are positive constants $c$ and $C$ such that $V_x/\bar{X} < c$, $\sigma_x/\mu_x < c$ and $0 < X_i < C < 1/2$. Then, for large $N$ and $n << N$,

$$\rho_z = -\frac{\sum_{i\in U} X_i^2 (Z_i - Y)^2}{\sum_{i\in U} X_i (Z_i - Y)^2}\left\{1 + O\left(\frac{1}{N}\right)\right\} + O\left(\frac{1}{N^2}\right). \qquad (10)$$

*Proof.* Because $\bar{X} = 1/N$, it follows from the above assumptions that the weighted mean $\mu_x$ $[= \Sigma X_i^2 = N(V_x^2 + \bar{X}^2)]$ is of order $1/N$ and hence, $\sigma_x = O(1/N)$. Because $(1-2X_i)^{-1} = 1 + 2X_i + O(X_i^2)$ for $0 < X_i < C < 1/2$, $\rho_z$ from (9) can be written for $N \to \infty$ as

$$\rho_z = -\sum_{i\in U}\frac{X_i^2}{\gamma}\left(\frac{Z_i - Y}{\sigma_z}\right)^2 + \frac{1}{\gamma} O\left(\sum_{i\in U} X_i^3\right),$$

where $\sum_{i\in U} X_i^3 = \sigma_x^2 + \mu_x^2 = O(N^{-2})$, and

$$\gamma = \frac{1}{2} + \frac{1}{2}\sum_{i\in U} X_i\{1 + 2X_i + O(X_i^2)\}$$

$$= 1 + \mu_x + O\left(\frac{1}{N^2}\right) = 1 + O\left(\frac{1}{N}\right),$$

from which (10) follows. This concludes the proof.

Substituting (10) into (3), we get

$$\text{var}(\hat{Y}_{\text{PPS}}) = \frac{\sigma_z^2}{n} - \frac{n-1}{n}\sum_{i\in U} X_i^2 (Z_i - Y)^2$$

$$= \frac{1}{n}\sum_{i\in U} X_i\{1 - (n-1)X_i\}(Z_i - Y)^2, \qquad (11)$$

which is also given by Hartley and Rao (1962). It is noteworthy that approximation (10) also follows directly from substituting the simple approximation $\pi_{ijAP} = n(n-1) X_i X_j$ into (4). Likewise, use of $\pi_{ijHR}$ leads to an expression almost similar to (9) and hence to (10). In addition, direct use of $\pi_{ijAP}$ in (1) or (2) for the SRS case with $X_i = X_j = 1/N$ may lead to errors of more than 100% for populations with $\bar{Y} = V_y^2$; see Knottnerus (2003, pages 274-6). Hence, (1) and (2) are more sensitive to small errors in the $\pi_{ij}$ than (3) and (4). Furthermore, note that when $n$ is so small that $|n\rho_z| << 1$, we may set $\rho_z = 0$ yielding the with-replacement variance formula of Hansen and Hurwitz (1943).

In order to estimate (3) using $\rho_z$, denote, as before, a randomly chosen observation from $s$ by $z_{s1}$. Then we have

$$\sigma_z^2 = \text{var}(z_{s1}) = \text{var}\{E(z_{s1}|s)\} + E\{\text{var}(z_{s1}|s)\}$$

$$= \text{var}(\bar{z}_s) + E\left(\frac{n-1}{n} s_z^2\right),$$

where

$$s_z^2 = \frac{1}{n-1}\sum_{i\in s} (Z_i - \bar{z}_s)^2.$$

Now from (3), it is seen that $s_z^2/(1-\rho_z)$ is an unbiased estimator for $\sigma_z^2$. When $\rho_z$ is very small, the term $(1-\rho_z)$ can be neglected. When $n$ is sufficiently large, the ratio $\rho_z$ from (9) can be estimated by

$$\hat{\rho}_{z9} = -\frac{\sum_{i\in s} X_i (Z_i - \bar{z}_s)^2 / \hat{\gamma}(1 - 2X_i)}{\sum_{i\in s} (Z_i - \bar{z}_s)^2},$$

where

$$\hat{\gamma} = \frac{1}{2} + \frac{1}{2n}\sum_{i\in s}\frac{1}{1-2X_i}.$$

Because $\hat{\gamma} \geq 1$ and $X_i \leq 1/n$, we have $\hat{\rho}_{z9} \geq -1/(n-2)$. For the bias of an estimated ratio when $n$ is small, see Cochran (1977, page 160).

In a similar manner $\rho_z$ from (10) can be estimated by

$$\hat{\rho}_{z10} = -\frac{\sum_{i\in s} X_i (Z_i - \bar{z}_s)^2}{\sum_{i\in s} (Z_i - \bar{z}_s)^2} \geq \frac{-1}{n} > \frac{-1}{n-1}.$$

Hence, replacing $\sigma_z^2$ and $\rho_z$ in (3) by $s_z^2/(1-\hat{\rho}_{z10})$ and $\hat{\rho}_{z10}$, respectively, leads to a nonnegative variance estimator

with probability 1. This also holds for $\hat{\rho}_{z9}$ when all $X_i \leq 1/(n+1)$. The estimator for $\text{var}(\hat{Y}_{\text{PPS}})$ thus obtained becomes

$$\hat{\text{var}}_{\rho}(\hat{Y}_{\text{PPS}}) = \frac{\{1 + (n-1)\hat{\rho}_{z9}\}s_z^2}{n(1-\hat{\rho}_{z9})}.$$

Moreover, for moderate values of $N$, estimator $\hat{\rho}_{z9}$ has probably better properties than $\hat{\rho}_{z10}$ because the $\pi_{ijK}$ underlying (9) satisfy exactly the second-order restrictions irrespective of the values of $n$ and $N$.

## 3. Efficiency of $\hat{Y}_{\text{PPS}}$ for large $n$ and $N$

### 3.1 Efficiency formulas

Because $X = 1$, the ratio estimator for $Y$ becomes

$$\hat{Y}_R = \frac{\bar{y}_s}{\bar{x}_s} = \frac{\sum_{i\in s} X_i Z_i}{\sum_{i\in s} X_i}.$$

For sufficiently large $n$ the commonly used approximation for its variance is

$$\text{var}(\hat{Y}_R) = \frac{N(N-n)}{n(N-1)} \sum_{i\in U} X_i^2 (Z_i - Y)^2. \qquad (12)$$

From (3) and (12) it can be seen that the efficiency of $\hat{Y}_{\text{PPS}}$ relative to $\hat{Y}_R$ can be written as

$$Eff_{P/R} = \frac{\text{var}(\hat{Y}_R)}{\text{var}(\hat{Y}_{\text{PPS}})} = \frac{(N-n)\sum_{i\in U} X_i^2 (Z_i - Y)^2}{\{1 + (n-1)\rho_z\}\sigma_z^2}, \qquad (13)$$

assuming $N/(N-1) \approx 1$. Combining (10) and (13) gives

$$Eff_{P/R} = \frac{-(N-n)\rho_z}{1 + (n-1)\rho_z}. \qquad (14)$$

Now suppose that the observations $Y_i$ satisfy the model:

$$Y_i = \mu X_i + \varepsilon_i, \qquad (15)$$

with $E(\varepsilon_i) = 0$, $E(\varepsilon_i^2) = \sigma^2 X_i^{\delta}$, and $E(\varepsilon_i \varepsilon_j) = 0$ $(i \neq j)$. Consequently, for the $Z_i$ we have $Z_i = \mu + u_i$ with $E(u_i) = 0$, $E(u_i^2) = \sigma^2 X_i^{\delta-2}$, and $E(u_i u_j) = 0$ $(i \neq j)$. According to Kott (1988), $\delta$ often lies between 1 and 2. See also Brewer (1963b). Brewer and Donadio (2003) showed that by assuming a model like (15), (7) and hence (10) and (14) hold when $n$ and $N$ are of the same order as $N \to \infty$. Furthermore, for sufficiently large $N$ we can replace $Y$ as well as the numerator and denominator in (10) by their model expectations. This yields

$$\rho_z = -\frac{\Sigma_{i\in U} X_i^{\delta}}{\Sigma_{i\in U} X_i^{\delta-1}}. \qquad (16)$$

In the next subsections we look more closely at the relationship between $\delta$ and the efficiency of $\hat{Y}_{\text{PPS}}$.

### 3.2 Efficiency of $\hat{Y}_{\text{PPS}}$ when $\delta = 2$

For $\delta = 2$, (16) gives $\rho_z = -\Sigma_{i\in U} X_i^2 = -\mu_x$, which can also be written as

$$\rho_z = -\frac{1}{N}(1 + CV_x^2), \qquad (17)$$

because

$$\frac{1}{N}\sum_{i\in U} X_i^2 = V_x^2 + \bar{X}^2 = \bar{X}^2(1 + CV_x^2),$$

where $\bar{X} = 1/N$ and $CV_x = V_x/\bar{X}$ is the coefficient of variation of the $X_i$. Substituting (17) into (14) gives

$$Eff_{P/R} = \frac{(N-n)(1+CV_x^2)}{N - (n-1)(1+CV_x^2)}.$$

Hence, for $\delta = 2$, the efficiency of the randomized PPS sample is high when the variability among the $X_i$ is high. When $CV_x = 0$, randomized PPS sampling amounts to SRS sampling and obviously, $Eff_{P/R} = 1$ assuming $(N - n + 1) \approx (N - n)$; note that this assumption holds when $N$ is sufficiently large and $n/N < f_0 < 1$.

Observe that substituting $n = n_{\text{PPS}}(1 + CV_x^2)$ into (12) leads to about the same outcome as (3) and (10) with $n_{\text{PPS}}$ instead of $n$. Hence, when $CV_x = 1.5$, randomized PPS sampling with sample size $n_{\text{PPS}} = 100$ is as efficient as the ratio estimator from an SRS sample of size $n_{\text{SRS}} = 325$. More generally, assuming that $(n-1)/n \approx 1$, it is seen from (3), (10), and (12) that a ratio estimator from an SRS sample of size $n_{\text{SRS}}$ is as efficient as a PPS sample of size $n_{\text{PPS}}$ when

$$n_{\text{SRS}} = -n_{\text{PPS}}\rho_z N. \qquad (18)$$

### 3.3 Efficiency of $\hat{Y}_{\text{PPS}}$ for $\delta < 1$ vs $\delta \geq 1$

Another special case is $\delta = 1$. From (16), $\rho_z = -1/N$ when $\delta = 1$. Subsequently, it follows from (14) that under model (15) $Eff_{P/R} = 1 + O(N^{-1})$, provided that $n/N < f_0 < 1$ as $N \to \infty$ irrespective of the value of $CV_x$. Furthermore, it can be shown that $Eff_{P/R}$ is an increasing function of $\delta$. This is proven below in Lemma 1. Hence, for $\delta < 1$ the randomized PPS estimator is less efficient than the ratio estimator, while for $\delta > 1$ the randomized PPS estimator is more efficient than the ratio estimator.

*Lemma* 1. Let $Eff_{P/R}$ and $\rho_z$ be defined by (14) and (16), respectively. If $V_x^2 > 0$, then $Eff_{P/R}$ is a monotonically increasing function of $\delta$.

*Proof.* Write $\rho_z$ from (16) as a weighted mean of the (negative) $X_i$

$$\rho_z = -u(\delta) = -\sum_{i \in U} w_i X_i,$$

where

$$w_i = \frac{X_i^{\delta-1}}{\sum_{i \in U} X_i^{\delta-1}} \quad \text{[Note that } \mu_x = u(2)\text{]}.$$

Let $X_i > X_j$ $(i \neq j)$, and define $h(\delta)$ as $w_i/w_j = (X_i/X_j)^{\delta-1}$. Since $h(\delta)$ is increasing in $\delta$, the weight of the larger $X_i$ is increasing compared to that of $X_j$ when $\delta$ is increasing. Hence, $u(\delta)$ is increasing and $\rho_z$ is decreasing in $\delta$. It suffices therefore to show that $Eff_{P/R}$ is decreasing in $\rho_z$. Writing (14) as

$$Eff_{P/R} = \frac{-(N-n)}{\rho_z^{-1} + (n-1)},$$

it is seen that $Eff_{P/R}$ is decreasing in $\rho_z$ indeed. This concludes the proof.

### 3.4 An alternative structure among the disturbances

Finally, suppose the variance of the disturbances in (15) is of the form:

$$\text{var}(\varepsilon_i) = c_1 X_i + c_2 X_i^2 \qquad (0 < c_1, c_2 \leq 1).$$

See Kott (1988). For this case we obtain in analogy with (16)

$$\rho_z = -\sum_{i \in U} \omega_i X_i,$$

where

$$\omega_i = \frac{1 + \phi X_i}{\sum_{i \in U}(1 + \phi X_i)}, \text{ and } \phi = c_2/c_1$$

when $\varphi = 0$, $\rho_z = -1/N$. Hence, when $c_2 = 0$, PPS sampling is only as efficient as the ordinary ratio estimator from SRS sampling. Along the same lines as the proof of Lemma 1, it can be shown that $\rho_z$ is decreasing in $\varphi$ while $Eff_{P/R}$ is increasing in $\varphi$ Hence, for this case the randomized PPS estimator is always more efficient than the ratio estimator when $c_2$ is positive.

## 4. An application to the Producer Price Index

The Producer Price Index (PPI) in The Netherlands is based on about 2,500 commodity price indexes organized by type of product. The price index for a specific commodity can be written as

$$Y = \sum_{i \in U} X_i Z_i,$$

where $Z_i$ is the price change for that commodity of establishment $i$ relative to the basic period while $X_i$ is the relative sales of that commodity by establishment $i$ in the basic period (recall $\Sigma X_i = 1$).

In the example given here, we examine the price changes of 70 establishments for the commodity *Basic Metal* in December of 2005 relative to December of 2004; see Table 1. We compare the variance of the ratio estimator from an SRS sample with the variance of the HT estimator from a randomized PPS sample when $n = 9$. Applying (12) to these data gives $\text{var}(\hat{Y}_R) = 101$. If the sample had been drawn with replacement the variance would have been 116. Applying (3) and (9) for a randomized PPS sample gives $\text{var}(\hat{Y}_{\text{PPS},\gamma}) = 29.9$. This outcome takes $\gamma$ into account and lies close to the result $V_{\text{PPS}}^{(sim)} = 29.2$ from a simulation experiment consisting of 80,000 randomized PPS samples of size $n = 9$ from the set of 70 establishments. Hence, $Eff_{P/R} = 3.5$. Because formula (12) for $\text{var}(\hat{Y}_R)$ is only asymptotically unbiased, we also carried out simulations evaluating the mean square error (MSE) and the bias of $\hat{Y}_R$ resulting in $\text{MSE}_R^{(sim)} = 108$ and a relatively small bias of 0.7. This confirms the conjecture that (12) gives an underestimation of the true variance; see Cochran (1977). Hence, for moderate samples the true value of $Eff_{P/R}$ might be somewhat higher than (14) suggests.

Furthermore, it is noteworthy that the simpler formula (10) for $\rho_z$ in combination with (3) gives almost the same result $\text{var}(\hat{Y}_{\text{PPS}}) = 30.7$ even though $N = 70$ is not very large. The with replacement PPS variance would have been 43.8. Hence, the variance reduction for randomized PPS sampling is more than 30% even though the sampling fraction $n/N$ is much smaller. According to (18), formula (12) with $n_{\text{SRS}} = 26$ gives about the same outcome as (3) with $n_{\text{PPS}} = 9$; note: $\rho_z = -0.042$. Hence, the sample sizes differ by a factor 2.9, which is more or less in line with the factor $(1 + CV_x^2) = 3.1$ from subsection 3.2. This should not be surprising because the price changes and their variability hardly depend on the sizes of the company. Fitting a double log regression

$$\ln(Z_i - Y)^2 = \alpha + \beta \ln X_i + v_i \qquad (19)$$

results in the estimate $\hat{\beta} = 0.07$ for the data in Table 1; units with $Z_i = Y$ should be omitted in the regression. The estimate $\hat{\beta} = 0.07$ corresponds with $\hat{\delta} = 2.07$ for the disturbances in (15) which explains the superiority of randomized PPS sampling for this type of data. Also for other commodities $\hat{\delta}$ often was about 2; see Enthoven (2007).

**Table 1**
**Price changes $(Z_i)$ and sizes $(X_i)$ of 70 establishments**

| $i$ | price change | size | $i$ | price change | size |
|---|---|---|---|---|---|
| 1 | -18.4% | 0.0608 | 36 | 34.8% | 0.0427 |
| 2 | -16.0% | 0.0784 | 37 | 13.1% | 0.0121 |
| 3 | 3.3% | 0.0762 | 38 | 31.7% | 0.0351 |
| 4 | 12.5% | 0.0100 | 39 | -24.8% | 0.0074 |
| 5 | 0.0% | 0.0029 | 40 | 55.3% | 0.0009 |
| 6 | 8.3% | 0.0006 | 41 | 40.5% | 0.0066 |
| 7 | -39.0% | 0.0182 | 42 | 34.6% | 0.0022 |
| 8 | -25.1% | 0.0020 | 43 | 1.7% | 0.0001 |
| 9 | 1.1% | 0.0040 | 44 | 0.0% | 0.0039 |
| 10 | 4.4% | 0.0066 | 45 | 3.9% | 0.0304 |
| 11 | -4.9% | 0.0039 | 46 | 25.4% | 0.0209 |
| 12 | -8.9% | 0.0070 | 47 | 25.6% | 0.0062 |
| 13 | -7.0% | 0.0148 | 48 | 0.0% | 0.0033 |
| 14 | -15.0% | 0.0108 | 49 | -0.3% | 0.0019 |
| 15 | -10.7% | 0.0087 | 50 | 66.6% | 0.0346 |
| 16 | -9.0% | 0.1079 | 51 | 0.0% | 0.0039 |
| 17 | -11.3% | 0.0247 | 52 | -2.9% | 0.0007 |
| 18 | 10.6% | 0.0024 | 53 | 15.8% | 0.0011 |
| 19 | -23.2% | 0.0001 | 54 | 0.0% | 0.0026 |
| 20 | -25.4% | 0.0001 | 55 | 0.0% | 0.0018 |
| 21 | -80.7% | 0.0002 | 56 | 11.6% | 0.0057 |
| 22 | 13.4% | 0.0005 | 57 | 0.0% | 0.0042 |
| 23 | -42.5% | 0.0010 | 58 | 0.0% | 0.0236 |
| 24 | -34.8% | 0.0014 | 59 | -1.5% | 0.0015 |
| 25 | -30.0% | 0.0126 | 60 | 0.0% | 0.0003 |
| 26 | 8.0% | 0.0530 | 61 | 11.7% | 0.0067 |
| 27 | 0.0% | 0.0208 | 62 | 0.0% | 0.0012 |
| 28 | 2.1% | 0.0119 | 63 | 0.8% | 0.0040 |
| 29 | 11.3% | 0.0208 | 64 | 2.0% | 0.0009 |
| 30 | 0.7% | 0.0322 | 65 | 2.3% | 0.0018 |
| 31 | 9.5% | 0.0447 | 66 | 4.7% | 0.0026 |
| 32 | 11.5% | 0.0018 | 67 | 0.9% | 0.0064 |
| 33 | 5.8% | 0.0174 | 68 | -1.0% | 0.0309 |
| 34 | -6.9% | 0.0197 | 69 | -0.5% | 0.0005 |
| 35 | 0.0% | 0.0124 | 70 | 0.0% | 0.0006 |

We conclude this section with a small example showing that randomized PPS is not *always* better than the ratio estimator. Although the data in Table 2 for a population of five units are artificial, a data pattern like this may occur in financial branches where very small financial companies may grow very fast with respect to certain financial variables. This high variability among growth rates of small companies results in a low value for $\delta$. For an SRS sample with $n = 2$ from the five units in Table 2 the variance of the ratio estimator is 211 according to (12); simulations give $\mathrm{MSE}_R^{(sim)} = 323$. This is much less than the variance of 557 found in a simulation consisting of 80,000 randomized PPS samples of size $n = 2$. Formula (3) in combination with (9) gives the same outcome: 557. This would also be the correct variance had sample been drawn according to Brewer (1963a) or Durbin (1967). Formula (11), based on (10), gives a slightly different value, 556.

Regression (19) with the data from Table 2 yields $\hat{\beta} = -3.0$, and hence $\hat{\delta} = -1.0$. In line with the findings of subsection 3.3 this low value $\hat{\delta} = -1.0$ explains why $\hat{Y}_{\mathrm{PPS}}$ is less efficient than $\hat{Y}_R$ in this example. Moreover, the ordinary direct estimator $N\bar{y}_s$ from an SRS sample has a variance of 356, which is even smaller here than the variance in randomized PPS sampling; $\bar{y}_s$ being the sample mean of the $Y_i$. Hence, for this type of data, the ratio estimator is the best option. Recall that the ratio estimator has a smaller variance than $N\bar{y}_s$ when $b > Y/2X$ where $b$ is the slope of a regression from $Y_i$ on $X_i$ and a constant $(i = 1, ..., N)$; see Knottnerus (2003, page 117). So the data $Y_i$ $(= X_i Z_i)$ in Table 2 certainly do not exhibit a flat trend.

**Table 2**
**Growth rates of assets $(Z_i)$ and sizes $(X_i)$ of 5 establishments**

| $i$ | growth rate | size |
|---|---|---|
| 1 | 200% | 0.0455 |
| 2 | 33% | 0.1364 |
| 3 | 75% | 0.1818 |
| 4 | 33% | 0.2727 |
| 5 | 62% | 0.3636 |

## 5. Summary

This paper compares the variance of the HT estimator $\hat{Y}_{\text{PPS}}$ from a randomized PPS sample with the variance of the classical ratio estimator $\hat{Y}_R$ from an SRS sample of the same size. In this comparison the sampling autocorrelation coefficient $\rho_z$ plays an important role.

When the data pattern of the variables $x$ and $z$ $(= y/x)$ is such that $\rho_z < -1/(N-1)$, it can be shown under mild conditions that $\hat{Y}_{\text{PPS}}$ is more efficient than $\hat{Y}_R$ for sufficiently large $n$ and $N$, provided that $X_i$ and $Z_i$ are uncorrelated. Under model (15) with $E(\varepsilon_i^2) = \sigma^2 X_i^{\delta}$ it holds that $\rho_z < -1/(N-1)$ when $\delta > 1$. Hence, for this type of data $\hat{Y}_{\text{PPS}}$ is to be preferred. Moreover, it emerges from (14) and (16) that for $\delta = 2$ the relative efficiency of PPS sampling compared to that of the ratio estimator is increasing when $CV_x$ is increasing. In addition, $\hat{Y}_R$ is to be preferred when the data correspond to a model with $\delta < 1$. These findings are confirmed empirically with a simulation study using two different data sets. When model (15) is not applicable, the relative efficiency of $\hat{Y}_{\text{PPS}}$ is given by (14) provided $n$ is large and $N$ is relatively larger. In practice the unknown $\rho_z$ in (14) is replaced by $\hat{\rho}_{z9}$. The fact that $n << N$ does not necessarily mean that the factor $(n-1)\rho_z$ in (3) is always negligible.

## Acknowledgements

The views expressed in the article are those of the author and do not necessarily reflect the policy of Statistics Netherlands. The author would like to thank Peter-Paul de Wolf, Sander Scholtus, the Associate Editor and two anonymous referees for their helpful suggestions and corrections.

## Appendix A

## A counterexample

Equations (5) and (7) cannot always be used for randomized PPS sampling when $n$ and $N$ are of the same order while $X_i$ and $Z_i$ are correlated. To see that, consider a population $U$ consisting of two groups $U_1$ and $U_2$ with means $\bar{Y}_1$ and $\bar{Y}_2$, respectively. Both stratum sizes are $N/2$. Let $s$ be a randomized PPS sample of size $n = 3N/4$ from the whole population $U$. Let the $X_i$ be such that

$$\pi_i = nX_i = \begin{cases} 1 & \text{if} \quad i \in U_1 \\ 0.5 & \text{if} \quad i \in U_2. \end{cases}$$

Obviously, group 1 does not contribute to the variance. The selected units in $s$ from $U_2$ constitute an ordinary SRS sample of size $N/4$. Hence, for randomized PPS sampling the correct variance formula in this example is

$$\text{var}(\hat{Y}_{\text{PPS}}) = \left(\frac{N}{2}\right)^2 \left(1 - \frac{1}{2}\right) \frac{S_{y2}^2}{N/4} = \frac{NS_{y2}^2}{2},$$

and

$$S_{y2}^2 = \frac{2}{N-2} \sum_{i \in U_2} (Y_{2i} - \bar{Y}_2)^2.$$

However, approximation (11) gives an entirely different, larger outcome unless $\bar{Y}_1 = 2\bar{Y}_2$.

## References

Brewer, K.R.W. (1963a). A model of systematic sampling with unequal probabilities. *Australian Journal of Statistics*, 5, 5-13.

Brewer, K.R.W. (1963b). Ratio estimation and finite population: Some results deductible from the assumption of an underlying stochastic process. *Australian Journal of Statistics*, 5, 93-105.

Brewer, K.R.W., and Donadio, M.E. (2003). The high entropy variance of the Horvitz-Thompson estimator. *Survey Methodology*, 29, 189-196.

Cochran, W.G. (1977). *Sampling Techniques.* New York: John Wiley & Sons, Inc.

Connor, W.S. (1966). An exact formula for the probability that two specified sampling units will occur in a sample drawn with unequal probabilities and without replacement. *Journal of the American Statistical Association*, 61, 384-390.

Cumberland, W.G., and Royall, R.M. (1981). Prediction models and unequal probability sampling. *Journal of the Royal Statistical Society*, B, 43, 353-367.

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

Durbin, J. (1967). Design of multi-stage surveys for the estimation of sampling errors. *Applied Statistics*, 16, 152-164.

Enthoven, L. (2007). *Cohort calculations* (in Dutch). Report MIC-2007-21, Statistics Netherlands, Voorburg.

Foreman, E.K., and Brewer, K.R.W. (1971). The efficient use of supplementary information in standard sampling procedures. *Journal of the Royal Statistical Society*, B, 33, 391-400.

Gabler, S. (1984). On unequal probability sampling: Sufficient conditions for the superiority of sampling without replacement. *Biometrika*, 71, 171-175.

Hansen, M.H., and Hurwitz, W.N. (1943). On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14, 333-362.

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

Hidiroglou, M.A., and Gray, G.B. (1980). Construction of joint probability of selection for systematic P.P.S. sampling. *Applied Statistics*, 29, 107-112.

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

Knottnerus, P. (2003). *Sample Survey Theory*: *Some Pythagorean Perspectives*. New York: Springer-Verlag.

Kott, P.S. (1988). Model-based finite population correction for the Horvitz-Thompson estimator. *Biometrika*, 75, 797-799.

Kott, P.S. (2005). A note on the Hartley-Rao variance estimator. *Journal of Official Statistics*, 21, 433-439.

Qualité, L. (2008). A comparison of conditional Poisson sampling versus unequal probability sampling with replacement. *Journal of Statistical Planning and Inference*, 138, 1428-1432.

Rosén, B. (1997). On sampling with probability proportional to size. *Journal of Statistical Planning and Inference*, 62, 159-191.

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

Sen, A.R. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 5, 119-127.

Stehman, S.V., and Overton, W.S. (1994). Comparison of variance estimators of the Horvitz-Thompson estimator for randomized variable probability systematic sampling. *Journal of the American Statistical Association*, 89, 30-43.

Thompson, M.E., and Wu, C. (2008). Simulation-based randomized systematic PPS sampling under substitution of units. *Survey Methodology*, 34, 3-10.

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

Survey Methodology, June 2011
Vol. 37, No. 1, pp. 103-108
Statistics Canada, Catalogue No. 12-001-X



# The use of estimating equations to perform a calibration on complex parameters

**Éric Lesage** [1]

Abstract

In the calibration method proposed by Deville and Särndal (1992), the calibration equations take only exact estimates of auxiliary variable totals into account. This article examines other parameters besides totals for calibration. Parameters that are considered complex include the ratio, median or variance of auxiliary variables.

Key Words: Calibration; Complex parameter; Estimating equation; Calibration weight.

## 1. Introduction

In survey statistics, two main approaches are used in the estimation phase: "model-assisted" estimators (such as the regression estimator or the ratio estimator) and calibration estimators (such as the raking ratio), proposed by Deville and Särndal (1992). The two approaches are somewhat similar, as shown by the regression estimator, which is the same as the calibration estimator with the $\chi^2$ distance ("linear" calibration method).

The purpose of this article is to expand the family of calibration estimators. With the current method, calibration can be performed on totals. The idea is to be able to take into account the calibration constraints of complex parameters or statistics such as a ratio, a median or a geometric mean. The reason for doing this is that auxiliary information may consist of a complex statistic rather than totals. For example, a ratio relative to the total population might be known, but not the total in the numerator or denominator.

The issue of complex parameters in calibrations has been discussed in the literature. Särndal (2007) reviewed a number of them, in particular the work of Harms and Duchesne (2006) on the calibration estimation of quantiles, and the work of Krapavickaite and Plikusas (2005) on calibration estimators of certain functions of totals.

The originality of the approach in this article is that it reduces calibration on a complex parameter to calibration on a total for a new *ad hoc* auxiliary variable. The advantage of this approach is that current calibration tools can be used and that there is no need to solve a complex optimization program.

In section 2 of the article, we review how the calibration method works, define calibration on complex parameters and describe simple cases in which calibration on a complex parameter can be reduced to calibration on a total. In section 3, we focus on parameters that can be defined as a solution to an estimating equation (Godambe and Thompson 1986). We introduce the concept of calibration on a complex parameter defined by an estimating equation and show that the resulting calibration equation can be replaced with an equation for calibration on a total.

## 2. A complex parameter defined as a function of totals

### 2.1 Review of calibration on totals

Let $U$ be a finite population of size $N$. The statistical units of the population are indexed by a label $k$, where $k \in \{1, \ldots, N\}$. A sample $s$ is selected using sample plan $p(s)$. Its size is denoted $n$ and may be random. Let $\pi_k$ be the probability that $k$ is included in sample $s$, and let $d_k = 1 / \pi_k$ be its sampling weight.

For any variable $z$ that takes the values $z_k$ for the units in $U$ indexed by $k$, the sum $t_z = \sum_{k \in U} z_k$ is referred to as the total of $z$ over $U$.

Let $y^{(1)}, \ldots, y^{(Q)}$ be $Q$ variables of interest, whose values are known only for sample $s$, and let $\theta_y$ be the parameter of interest that is a function of the totals $t_{y^{(1)}}, \ldots, t_{y^{(Q)}}$:

$$\theta_{\mathbf{y}} = f(t_{y^{(1)}}, \ldots, t_{y^{(Q)}}).$$

The estimator of $\theta_{\mathbf{y}}$ is

$$\hat{\theta}_{\mathbf{y},\pi} = f(\hat{t}_{y^{(1)},\pi}, \ldots, \hat{t}_{y^{(Q)},\pi}).$$

It is simply the function $f(\cdot, \ldots, \cdot)$ with totals $t_{y^{(q)}}$ replaced by their Horvitz-Thompson estimator $\hat{t}_{y^{(q)},\pi} = \sum_{k \in s} d_k y_k^{(q)}$ (Särndal, Swensson and Wretman 1992). This estimator can be described as a substitution estimator.

Let $x^{(1)}, \ldots, x^{(P)}$ be $P$ auxiliary variables known on $s$, and let $t_{x^{(1)}}, \ldots, t_{x^{(P)}}$ be the totals on $U$ for those auxiliary variables, also known. For an individual $k$, the vector of values taken by the auxiliary variables on $k$ is denoted $\mathbf{x}_k' = (x_k^{(1)}, \ldots, x_k^{(P)})$.

The calibration estimator of $\theta_{\mathbf{y}}$ is

1. Éric Lesage, CREST(ENSAI).and IRMAR(UEB). Ker Lann Campus, F-35172 BRUZ, France. E-mail: eric.lesage@ensai.fr.

$$\hat{\theta}_{\mathbf{y},\mathrm{CAL}} = f(\hat{t}_{y^{(1)},\mathrm{CAL}}, \ldots, \hat{t}_{y^{(Q)},\mathrm{CAL}})$$

with $\hat{t}_{y^{(q)},\mathrm{CAL}} = \sum_{k\in s} w_k\, y_k^{(q)}$, and a series of weights $\{w_k\}_{(k\in s)}$, known as calibration weights (which should be denoted $w_k(s)$, since they depend on the sampling), obtained by solving the following optimization program:

$$\min_{\{w_k\}_{(k\in s)}} \sum_{k\in s} d(w_k, d_k)$$

under constraints

$$\begin{cases} \hat{t}_{x^{(1)},\mathrm{CAL}} = t_{x^{(1)}} \\ \ldots \\ \hat{t}_{x^{(P)},\mathrm{CAL}} = t_{x^{(P)}} \end{cases}$$

$d(\cdot,\cdot)$ is a pseudo-distance, *i.e.*, a function that measures the difference between the calibration weight and the sampling weight (unlike a difference, a pseudo-distance is not necessarily symmetrical on its two arguments). The program is solved with a Lagrangian. When the distance used is the $\chi^2$ distance (*i.e.*, $d(w_k, d_k) = (1/2)(w_k - d_k)^2/d_k$), the solution is $w_k = d_k(1 + \mathbf{x}_k'\lambda)$ (where $\lambda$ is a $P$-vector of Lagrange multipliers).

### 2.2 Calibration on a complex parameter $\eta_{\mathbf{x}}$

*Definition* 1: *Let* $x^{(1)}, \ldots, x^{(P)}$ *be P auxiliary variables known on s, and let* $\eta_{\mathbf{x}} = g(t_{x^{(1)}}, \ldots, t_{x^{(P)}})$ *be a complex parameter, a function of the totals of those auxiliary variables, also known.*

*In the case of calibration on the complex parameter* $\eta_{\mathbf{x}}$, *the calibration weights are obtained by solving the following optimization program*:

$$\min_{\{w_k\}_{(k\in s)}} \sum_{k\in s} d(w_k, d_k)$$

*under constraints*

$$\hat{\eta}_{\mathbf{x},\mathrm{CAL}} = g(\hat{t}_{x^{(1)},\mathrm{CAL}}, \ldots, \hat{t}_{x^{(P)},\mathrm{CAL}}) = \eta_{\mathbf{x}}.$$

The totals $t_{x^{(q)}}$ do not have to be known, but the complex parameter $\eta_{\mathbf{x}}$ does.

Consider the example of the ratio

$$R_{\mathbf{x}} = \frac{t_{x^{(1)}}}{t_{x^{(2)}}} = \frac{\sum_{k\in U} x_k^{(1)}}{\sum_{k\in U} x_k^{(2)}}.$$

The calibration estimator of $R_{\mathbf{x}}$ is of the form

$$\hat{R}_{\mathbf{x},\mathrm{CAL}} = \frac{\sum_{k\in s} w_k x_k^{(1)}}{\sum_{k\in s} w_k x_k^{(2)}}.$$

The calibration equation in the case of calibration on a ratio is

$$\hat{R}_{\mathbf{x},\mathrm{CAL}} = \frac{\sum_{k\in s} w_k x_k^{(1)}}{\sum_{k\in s} w_k x_k^{(2)}} = R_{\mathbf{x}}$$

$R_{\mathbf{x}}$ is known auxiliary information, as the total of the auxiliary variables usually is. This scenario may occur when we have proportions that are well known and stable over time, for example, but the specific totals in the numerator and denominator are not known.

We described the case of calibration on a single complex parameter, but it is clearly a simple matter to calibrate on more than one complex parameter. In that case, there are as many constraints as calibration parameters.

### 2.3 Simple cases where calibration on a complex parameter can be reduced to calibration on a total

It is not easy to determine from the outset whether an equation for calibration on a complex parameter can be written in the form of an equation for calibration on a total. In other words, it is not always a trivial matter to find a "new" auxiliary variable $z$, associated with the complex parameter, on whose total we can calibrate.

For example, that is quite straightforward for all moments of an auxiliary variable $x$ (it is assumed that under the sampling plan, the population size $N$ can be estimated exactly). If $\mu_{x^m} = N^{-1}\sum_{k\in U} x_k^m$ is auxiliary information, we can simply take $z_k = x_k^m / N$ and calibrate on $\mu_{x^m}$: $\sum_{k\in s} w_k\, x_k^m/N = \mu_{x^m}$.

If we want to calibrate on the variance and the mean of variable $x$ with $\mu_x$ and $\sigma_x^2$ as auxiliary information, we can use the two new auxiliary variables

$$z_k^{(1)} = \frac{x_k}{N}$$

and

$$z_k^{(2)} = \frac{(x_k - \mu_x)^2}{N}.$$

On the other hand, if we do not know $\mu_x$, but we have $\sigma_x^2$ in the auxiliary information and we want to calibrate on that variance, things become more complicated. We can see this if we write the substitution estimator of $\sigma_x^2$ (where the sampling plan allows the population size $N$ to be estimated exactly):

$$\hat{\sigma}^2_{x,\mathrm{CAL}} = \frac{1}{N}\sum_{k\in s} w_k \left( x_k - \left( \frac{\sum_{l\in s} w_l x_l}{N} \right) \right)^2.$$

Finding a new auxiliary variable $z$ is not straightforward, since the initial calibration equation is not linear relative to the weight vector. We will return to the variance case in section 3.3 below.

*Ratio example*

*Proposition* 1: *Calibration on a ratio is equivalent to calibration on the total of the new auxiliary variable*: $z_k = x_k^{(1)} - R_{\mathbf{x}} x_k^{(2)}$.

*The calibration equation is written*

$$\hat{t}_{z,\mathrm{CAL}} = t_z = 0.$$

*Proof*:

$$\hat{t}_{z,\mathrm{CAL}} = t_z$$

$$\Leftrightarrow \sum_{k\in s} w_k (x_k^{(1)} - R_{\mathbf{x}} x_k^{(2)}) = \sum_{k\in U} (x_k^{(1)} - R_{\mathbf{x}} x_k^{(2)})$$

$$\Leftrightarrow \hat{t}_{x^{(1)},\mathrm{CAL}} - R_{\mathbf{x}}\, \hat{t}_{x^{(2)},\mathrm{CAL}} = t_{x^{(1)},\mathrm{CAL}} - R_{\mathbf{x}}\, t_{x^{(2)},\mathrm{CAL}} = 0$$

$$\Leftrightarrow \frac{\hat{t}_{x^{(1)},\mathrm{CAL}}}{\hat{t}_{x^{(2)},\mathrm{CAL}}} = R_{\mathbf{x}}$$

*i.e.*, $\hat{R}_{\mathbf{x},\mathrm{CAL}} = R_{\mathbf{x}}$.

*Function of a ratio of linear combinations of totals*

Let $\eta_{\mathbf{x}}$ be a complex parameter that is a bijective function of a ratio of linear combinations of totals:

$$\eta_{\mathbf{x}} = h\left(\frac{\alpha' \cdot \mathbf{t}_{\mathbf{x}}}{\beta' \cdot \mathbf{t}_{\mathbf{x}}}\right) \qquad (1)$$

with $\alpha' = (\alpha_1, ..., \alpha_P)$ and $\beta' = (\beta_1, ..., \beta_P)$ being vectors of real coefficients of size $P$, and $\mathbf{t}'_{\mathbf{x}} = (t_{x^{(1)}}, \ldots, t_{x^{(P)}})$.

*Proposition* 2: *Performing a calibration on complex parameter* $\eta_{\mathbf{x}}$ *defined by function* (1) *is equivalent to calibrating on the total of the new auxiliary variable*:

$$z_k = (\alpha' - h^{-1}(\eta_{\mathbf{x}})\beta') \cdot \mathbf{x_k}$$

*with calibration equation*

$$\hat{t}_{z,\mathrm{CAL}} = \sum_{k\in s} w_k z_k = t_z = 0.$$

*Proof*:

$$\hat{\eta}_{\mathbf{x},\mathrm{CAL}} = \eta_{\mathbf{x}} \Leftrightarrow h\left(\frac{\alpha' \cdot \hat{\mathbf{t}}_{\mathbf{x,CAL}}}{\beta' \cdot \hat{\mathbf{t}}_{\mathbf{x,CAL}}}\right) = \eta_{\mathbf{x}}$$

$$\Leftrightarrow \frac{\alpha' \cdot \hat{\mathbf{t}}_{\mathbf{x,CAL}}}{\beta' \cdot \hat{\mathbf{t}}_{\mathbf{x,CAL}}} = h^{-1}(\eta_{\mathbf{x}})$$

$$\Leftrightarrow (\alpha' - h^{-1}(\eta_{\mathbf{x}})\, \beta') \cdot \hat{\mathbf{t}}_{\mathbf{x,CAL}} = 0$$

$$\Leftrightarrow \sum_{k\in s} w_k (\alpha' - h^{-1}(\eta_{\mathbf{x}})\, \beta') \cdot \mathbf{x_k} = 0.$$

Consider the example of the geometric mean:

$$\mu_{\mathrm{Geo},\mathbf{x}} = \left(\prod_{k\in U} x_k\right)^{1/N}.$$

This expression can be rewritten as

$$\mu_{\mathrm{Geo},\mathbf{x}} = \exp\left(\frac{\sum_{k\in U} \ln(x_k)}{\sum_{k\in U} 1}\right).$$

We denote $\mathbf{x}'_k = (x_k^{(1)}, x_k^{(2)}) = (\ln(x_k), 1)$, $\alpha' = (1, 0)$, $\beta' = (0, 1)$ and $h^{-1}(u) = \exp^{-1}(u) = \ln(u)$.

Hence, the new auxiliary variable is

$$z_k = \ln(x_k) - \ln(\mu_{\mathrm{Geo},\mathbf{x}}) \cdot 1.$$

We will see later in the article that the estimating equations method provides another approach to displaying the new auxiliary variable(s) **z**.

## 3. Parameter defined by an estimating equation

### 3.1 Estimating with an estimating equation

Certain parameters $\theta_y$ are defined, or can be defined, as the solution to an implicit function known as the *estimating equation on U* (Godambe and Thompson 1986), *i.e.*:

$$\sum_{k\in U} \Phi(\theta_{\mathbf{y}}, \mathbf{y}_k) = 0$$

with $\mathbf{y}'_k = (y_k^{(1)}, ..., y_k^{(Q)})$ being the vector of values taken by the variables of interest for individual *k*.

In this context, an estimator of $\theta_{\mathbf{y}}$ is defined for sample *s,* denoted $\hat{\theta}_{\mathbf{y},ee,\pi}$, which is the solution of the *estimating equation on s* (see in particular Hidiroglou, Rao and Yung 2002):

$$\sum_{k\in s} d_k \Phi(\hat{\theta}_{\mathbf{y},ee,\pi}, \mathbf{y}_k) = 0.$$

**Table 1**
**Examples of parameters defined by estimating equations on *U***

| Parameter | $\Phi(\theta_{\mathbf{y}}, \mathbf{y}_k)$ | Estimating equation on *U* |
|---|---|---|
| mean $\mu$ | $(\mathbf{y}_k - \mu)$ | $\sum_{k\in U} (y_k - \mu) = 0$ |
| ratio $R = \mu_1 / \mu_2$ | $(y_k^{(1)} - R\, y_k^{(2)})$ | $\sum_{k\in U} (y_k^{(1)} - R y_k^{(2)}) = 0$ |
| median *m* | $(1_{y_k \le m} - 1/2)$ | $\sum_{k\in U} (1_{y_k \le m} - 1/2) = 0$ |

Consider also the example of the coefficient of a logistic regression. Let $y^{(1)}$ be a dichotomous variable that takes the values 0 and 1 on *U*, and let $y^{(2)}$ be a quantitative variable. The value $y_k^{(1)}$ taken by $y^{(1)}$ for unit *k* is assumed to be an instance of the random variable $Y_k^{(1)}$, which has a Bernoulli distribution

$$\mathfrak{B}\left(1,\ p_k = \frac{1}{1 + \exp(-\beta_0 y_k^{(2)})}\right).$$

We have limited the number of parameters to one, but it would be just as simple to consider the multidimensional case. However, we should provide a definition of the estimating equations that take the case of the vector parameters into account.

The parameter of interest to us is the estimator of $\beta_0$, denoted $\beta$, calculated on the finite population by the maximum likelihood method. The estimating equation of $\beta$ on $U$ will be the maximum likelihood equation. The log-likelihood in the case of Bernoulli variables is

$$\mathbf{L}(\beta) = \sum_{k\in U} y_k^{(1)} \ln(p_k) + \sum_{k\in U} (1 - y_k^{(1)}) \ln(1 - p_k).$$

It is easy to derive the estimating equation of $\beta$ on $U$:

$$\sum_{k\in U} y_k^{(2)} \left( y_k^{(1)} - \frac{1}{1+\exp(-\beta y_k^{(2)})} \right) = 0.$$

The estimating equation on $s$ which defines the estimator $\hat{\beta}_{ee,\pi}$ on the basis of the sampling weights is

$$\sum_{k\in s} d_k y_k^{(2)} \left( y_k^{(1)} - \frac{1}{1+\exp(-\hat{\beta}_{ee,\pi} y_k^{(2)})} \right) = 0.$$

The estimating equation is not linear in the parameter; $\hat{\beta}_{ee,\pi}$ cannot be expressed as a simple function of the observations.

The logistic regression example is very interesting because it shows that we do not need to know $\hat{\beta}_{ee,\pi}$ to perform the calibration. We will see in the next subsection that we only need to know the generic term of the estimating equation on

$$U,\ \Phi(\beta, \mathbf{y}_k) = y_k^{(2)} \left( y_k^{(1)} - \frac{1}{1+\exp(-\beta y_k^{(2)})} \right),$$

for all $k \in s$.

### 3.2 Calibration in the case of parameters defined by estimating equations

Let $\mathbf{x}'_k = (x^{(1)}, \ldots, x^{(P)})$ be the vector of $P$ known auxiliary variables on $s$, and let $\eta_{\mathbf{x}}$ be a complex parameter, also known, defined by the estimating equation

$$\sum_{k\in U} \Psi(\eta_{\mathbf{x}}, \mathbf{x}_k) = 0.$$

*Definition 2*: *In the case of calibration on the complex parameter* $\eta_{\mathbf{x}}$, *the calibration weights are obtained by solving the following optimization program*:

$$\min_{\{w_k\}_{(k\in s)}} \sum_{k\in s} d(w_k, d_k)$$

*under constraints*

$$\sum_{k\in s} w_k \Psi(\eta_{\mathbf{x}}, \mathbf{x}_k) = 0.$$

*Proposition* 3: *Calibration on a complex parameter* $\eta_{\mathbf{x}}$, *defined by an estimating equation*, *is equivalent to a calibration on the total of the new auxiliary variable*: $z_k = \Psi(\eta_{\mathbf{x}}, \mathbf{x}_k)$, *with the calibration constraint* $\sum_{k\in s} w_k z_k = 0$.

*Definition* 3: *A calibration estimator of the parameter of interest* $\theta_{\mathbf{y}}$, *denoted* $\hat{\theta}_{\mathbf{y},ee,\mathrm{CAL}}$, *is a solution to the estimating equation on* $s$ *weighted by the calibration weights* $\{w_k\}_{(k\in s)}$:

$$\sum_{k\in s} w_k \Phi(\hat{\theta}_{\mathbf{y},ee,\mathrm{CAL}}, \mathbf{y}_k) = 0.$$

In most cases, the solution to the estimating equation is unique. The median is an example of a parameter for which there may be more than one solution. In this case, the infimum is often used as an estimator.

*Proposition* 4: *If there is only one solution to the equation* $\sum_{k\in s} w_k \Psi(\hat{\eta}_{\mathbf{x},ee,\mathrm{CAL}}, \mathbf{x}_k) = 0$, *then*

$$\hat{\eta}_{\mathbf{x},ee,\mathrm{CAL}} = \eta_{\mathbf{x}}.$$

*Proof*: $\eta_{\mathbf{x}}$ *is a solution to the estimating equation that defines* $\hat{\eta}_{\mathbf{x},ee,\mathrm{CAL}}$. *Since there is a unique solution*, *we have* $\hat{\eta}_{\mathbf{x},ee,\mathrm{CAL}} = \eta_{\mathbf{x}}$.

### 3.3 Calibration on a variance

In this section, we examine calibration on variance $\sigma_x^2$, which is a more complicated complex parameter than those discussed above. We will show that when the variance is the only auxiliary information we have, we can perform an approximate calibration that produces calibration weights that have better properties than the sampling weights.

Back to the variance case. The mean $\mu_x$ and the variance $\sigma_x^2$ on $U$ of auxiliary variable $x$ can be defined by two estimating equations on $U$:

$$\begin{cases} \sum_{k\in U} (x_k - \mu_x) = 0 & (2) \\ \sum_{k\in U} ((x_k - \mu_x)^2 - \sigma_x^2) = 0. & (3) \end{cases}$$

If we know the two parameters, calibrating on them is easy, since we merely have to calibrate on the totals of the two new auxiliary variables $z^{(1)} = x - \mu_x$ and $z^{(2)} = (x - \mu_x)^2 - \sigma_x^2$.

On the other hand, if we consider the textbook case where the mean $\mu_x$ is not known, the parameter $\sigma_x^2$ cannot be defined by a unique estimating equation. If we replace $\mu_x$ with its explicit definition

$$\mu_x = \frac{\sum_{l\in U} x_l}{\sum_{j\in U} 1}$$

in equation (3), we obtain the equation

$$\sum_{k\in U}\left(\left(x_k-\frac{\sum_{l\in U}x_l}{\sum_{j\in U}1}\right)^2-\sigma_x^2\right)=0,$$

which cannot be written in the form of an estimating equation: $\sum_{k\in U}\Psi(\sigma_x^2,x_k)=0$.

$\mu_x$ thus becomes a nuisance parameter (Binder 1991). To overcome this difficulty, we can replace it in equation (3) with its substitution estimator: $\hat{\mu}_{x,\pi}=\hat{t}_{x,\pi}/\hat{N}_\pi$, with $\hat{N}_\pi=\sum_{k\in s}d_k 1$ being the Horvitz-Thompson estimator of the size of population $U$. This leads to the "approximate" calibration equation

$$\sum_{k\in s}w_k\left(\left(x_k-\frac{\hat{t}_{x,\pi}}{\hat{N}_\pi}\right)^2-\sigma_x^2\right)=0. \qquad (4)$$

*Proposition 5*: *With estimating equation* (4), *calibration on the variance is not perfect*, *and we have*

$$\hat{\sigma}^2_{x,ee,\mathrm{CAL}}=\sigma_x^2-\left(\frac{\hat{t}_{x,\pi}}{\hat{N}_\pi}-\frac{\hat{t}_{x,\mathrm{CAL}}}{\hat{N}_{\mathrm{CAL}}}\right)^2. \qquad (5)$$

*Proof*:

- The "approximate" calibration equation is equation (4).
- The definition of the parameters' calibration estimators:

$$\begin{cases}\sum_{k\in s}w_k(x_k-\hat{\mu}_{x,ee,\mathrm{CAL}})=0\\ \sum_{k\in s}w_k((x_k-\hat{\mu}_{x,ee,\mathrm{CAL}})^2-\hat{\sigma}^2_{x,ee,\mathrm{CAL}})=0.\end{cases}$$

  This can be rewritten

$$\begin{cases}\hat{\mu}_{x,ee,\mathrm{CAL}}=\dfrac{\sum_{k\in s}w_kx_k}{\sum_{k\in s}w_k}=\dfrac{\hat{t}_{x,\mathrm{CAL}}}{\hat{N}_{\mathrm{CAL}}}\\ \sum_{k\in s}w_k\left(\left(x_k-\dfrac{\hat{t}_{x,CAL}}{\hat{N}_{\mathrm{CAL}}}\right)^2-\hat{\sigma}^2_{x,ee,\mathrm{CAL}}\right)=0.\end{cases}$$

- If we subtract the second estimating equation from the approximate calibration equation, we get

$$\sum_{k\in s}w_k\left(\left(x_k-\frac{\hat{t}_{x,\pi}}{\hat{N}}\right)^2-\left(x_k-\frac{\hat{t}_{x,\mathrm{CAL}}}{\hat{N}_{\mathrm{CAL}}}\right)^2-\sigma_x^2+\hat{\sigma}^2_{x,ee,\mathrm{CAL}}\right)=0.$$

  Using the identity $a^2-b^2=(a-b)(a+b)$, we have

$$\sum_{k\in s}w_k\left(\left(\frac{\hat{t}_{x,\mathrm{CAL}}}{\hat{N}_{\mathrm{CAL}}}-\frac{\hat{t}_{x,\pi}}{\hat{N}_\pi}\right)\left(2x_k-\frac{\hat{t}_{x,\pi}}{\hat{N}_\pi}-\frac{\hat{t}_{x,\mathrm{CAL}}}{\hat{N}_{\mathrm{CAL}}}\right)\right)$$
$$-\hat{N}_{\mathrm{CAL}}(\sigma_x^2-\hat{\sigma}^2_{x,ee,\mathrm{CAL}})=0$$

$$\left(\frac{\hat{t}_{x,\mathrm{CAL}}}{\hat{N}_{\mathrm{CAL}}}-\frac{\hat{t}_{x,\pi}}{\hat{N}_\pi}\right)\sum_{k\in s}w_k\left(2x_k-\frac{\hat{t}_{x,\pi}}{\hat{N}_\pi}-\frac{\hat{t}_{x,\mathrm{CAL}}}{\hat{N}_{\mathrm{CAL}}}\right)$$
$$-\hat{N}_{\mathrm{CAL}}(\sigma_x^2-\hat{\sigma}^2_{x,ee,\mathrm{CAL}})=0$$

$$\left(\frac{\hat{t}_{x,\mathrm{CAL}}}{\hat{N}_{\mathrm{CAL}}}-\frac{\hat{t}_{x,\pi}}{\hat{N}_\pi}\right)\left(2\hat{t}_{x,\mathrm{CAL}}-\frac{\hat{t}_{x,\pi}}{\hat{N}_\pi}\hat{N}_{\mathrm{CAL}}-\hat{t}_{x,\mathrm{CAL}}\right)$$
$$-\hat{N}_{\mathrm{CAL}}(\sigma_x^2-\hat{\sigma}^2_{x,ee,\mathrm{CAL}})=0$$

$$\hat{N}_{\mathrm{CAL}}\left(\frac{\hat{t}_{x,\mathrm{CAL}}}{\hat{N}_{\mathrm{CAL}}}-\frac{\hat{t}_{x,\pi}}{\hat{N}_\pi}\right)^2-\hat{N}_{\mathrm{CAL}}(\sigma_x^2-\hat{\sigma}^2_{x,ee,\mathrm{CAL}})=0.$$

  This is the same as the expression for $\hat{\sigma}^2_{x,ee,\mathrm{CAL}}$ in equation (5).

This result is interesting because, without an exact calibration, we have a calibration estimator of $\sigma_x^2$ that is asymptotically more precise than the substitution estimator $\hat{\sigma}^2_{x,\pi}$. That is, if we resort to the asymptotic framework typically used in surveys and employ linearization of complex estimators (Deville 1999), we have

$$\hat{\sigma}^2_{x,\pi}-\sigma_x^2=O_p\left(\frac{1}{\sqrt{n}}\right)$$

and

$$(\hat{\sigma}^2_{x,ee,\mathrm{CAL}}-\sigma_x^2)^{1/2}=\left(\frac{\hat{t}_{x,\pi}}{\hat{N}_\pi}-\frac{\hat{t}_{x,\mathrm{CAL}}}{\hat{N}_{\mathrm{CAL}}}\right)=O_p\left(\frac{1}{\sqrt{n}}\right).$$

This yields

$$\hat{\sigma}^2_{x,ee,\mathrm{CAL}}-\sigma_x^2=O_p\left(\frac{1}{n}\right).$$

## 4. Conclusion

In this article, we presented a simple method of performing a calibration in cases where the auxiliary information takes the form of a complex parameter. That method is based on the concept of the estimating equation. Its major advantage is that it can be used with current calibration software.

In future research, it would be interesting to determine the practical cases in which the use of complex parameters in the calibration improves the precision of the parameters of interest.

## Acknowledgements

The author wishes to thank the journal's associate editor, the reviewers, Guillaume Chauvet, François Coquet and Jean-Claude Deville for their constructive comments on the preliminary versions of this paper.

## References

Binder, D.A. (1991). Use of estimating functions for interval estimation from complex surveys. *Proceedings of the survey research methods section*, American Statistical Association, 34-42.

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

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

Godambe, V.P., and Thompson, M.E. (1986). Parameters of superpopulation and survey population: Their relationship and estimation. *International Statistical Review*, 54, 127-138.

Harms, T., and Duchesne, P. (2006). On calibration estimation for quantiles. *Survey Methodology*, 32, 37-52.

Hidiroglou, M., Rao, J.N.K. and Yung, W. (2002). Estimating equations for the analysis of survey data using poststratification information. *Sankhyā*, 64, 2, 364-378.

Krapavickaite, D., and Plikusas, A. (2005). Estimation of ratio in finite population. *Informatica*, 16, 347-364.

Plikusas, A. (2006). Non-linear calibration. Recueil du Colloque sur les méthodes de sondage, Venspils, Latvia. Riga: Central Statistical Bureau of Latvia.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New-York: Springer-Verlag, 162-163.

Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, 33, 99-119.

# 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 26, No. 4, 2010**



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 27, No. 1, 2011**



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 38, No. 4, December/décembre 2010**

CONTENTS TABLE DES MATIÈRES

**Volume 39, No. 1, March/mars 2011**