Catalogue no. 12-001-X

Survey Methodology

SURVEY
METHODOLOGY

Catalogue No. 12-001-XPB

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 2005
•
VOLUME 31
•
NUMBER 1

Statistics Canada Statistique Canada

Canada

December 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. 12-000-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 Finance R.H. Coats Bldg., 6th Floor 150 Tunney's Pasture Driveway Ottawa, Ontario K1A 0T6
• In person from authorized agents and bookstores.	

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

Standards of service to the public

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

Statistics Canada

Business Survey Methods Division

Survey Methodology

December 2011

Published by authority of the Minister responsible for Statistics Canada

December 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 (n^{o} 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 2, December 2011

Contents

Waksberg Invited Paper Series

Regular Papers

Special Section of the U.S. Census Bureau

Introduction

Special Section Papers

Discussion

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.
∞

Waksberg Invited Paper Series

The journal *Survey Methodology* has established an annual invited paper series in honour of Joseph Waksberg, who has made many important contributions to survey methodology. Each year a prominent survey researcher is chosen to author an article as part of the Waksberg Invited Paper Series. The paper reviews the development and current state of a significant topic within the field of survey methodology, and reflects the mixture of theory and practice that characterized Waksberg's work.

Please see the announcements at the end of the Journal for information about the nomination and selection process of the 2012 Waksberg Award.

This issue of *Survey Methodology* opens with the tenth paper of the Waksberg Invited Paper Series. The editorial board would like to thank the members of the selection committee Daniel Kasprzyk (Chair), Elisabeth A. Martin, Mary E. Thompson and Wayne Fuller for having selected Danny Pfeffermann as the author of this year's Waksberg Award paper.

2011 Waksberg Invited Paper

Author: Danny Pfeffermann

Danny Pfeffermann is Professor of statistics at the Hebrew University of Jerusalem, Israel, and at Southampton Statistical Sciences Research Institute (S3RI), University of Southampton, UK. For the past 15 years he is also a consultant for the US Bureau of Labor Statistics. His main research areas are analytic inference from complex sample surveys, seasonal adjustment and trend estimation, small area estimation, and more recently, observational studies and nonresponse. Danny served for two years as the president of the Israel Statistical Association and is the president elect of the International Association of Survey Statisticians (IASS). He is co-editor of the new two-volume handbook in Statistics on "Sample Surveys".

Waksberg Invited Paper Series

Preface from the author

It is a great honour to receive the award named after Joe Waksberg. I am old enough to have had the fortune of meeting Joe on several occasions, the last time being a whole day of professional meetings at Westat, discussing nothing else but my own modest contributions to survey sampling. What I remember from these meetings is Joe's brilliance, profound knowledge and sharp intellect, even at his very advanced age. I would be lying if I say that I was able to answer all his critical questions.

I feel even more honoured and privileged when I look at the list of all the eminent survey statisticians who received the award before me. While I am still trying to convince myself that I deserve being on that list, I am overwhelmed by all the sincere congratulations and good words from colleagues around the world and during the symposium. What can I say, I am very proud and grateful.

On this occasion, I would like to commemorate also one of the founders and the long serving editor of *Survey Methodology*, the late M.P. Singh. In 1993 I published a paper in the *International Statistical Review* entitled "The role of sampling weights when modeling survey data". This paper was well received and when I met M.P. a couple of years later, he sort of complained to me for not publishing the paper in *Survey Methodology*. Not having a convincing answer, I promised M.P. that one day I would write another paper on this topic and submit it to Survey Methodology. I feel that with the present paper I have kept my promise to M.P. Singh.

Danny Pfeffermann

Survey Methodology, December 2011
Vol. 37, No. 2, pp. 115-136
Statistics Canada, Catalogue No. 12-001-X

Modelling of complex survey data: Why model? Why is it a problem? How can we approach it?

Danny Pfeffermann [1]

Abstract

This article attempts to answer the three questions appearing in the title. It starts by discussing unique features of complex survey data not shared by other data sets, which require special attention but suggest a large variety of diverse inference procedures. Next a large number of different approaches proposed in the literature for handling these features are reviewed with discussion on their merits and limitations. The approaches differ in the conditions underlying their use, additional data required for their application, goodness of fit testing, the inference objectives that they accommodate, statistical efficiency, computational demands, and the skills required from analysts fitting the model. The last part of the paper presents simulation results, which compare the approaches when estimating linear regression coefficients from a stratified sample in terms of bias, variance, and coverage rates. It concludes with a short discussion of pending issues.

Key Words: Informative sampling; NMAR nonresponse; Likelihood-based methods; Probability weighting; Randomization distribution; Sample model.

1. Introduction

Survey data are frequently used for analytic inference on statistical models, which are assumed to hold for the population from which the sample is taken. Familiar examples include the estimation of income elasticities from household surveys, the analysis of labour market dynamics from labour force surveys, comparisons of pupils' achievements from educational surveys and the search for causal relationships between risk factors and disease prevalence from health surveys. An important common feature to all these examples is that interest lies in the structure of the models being estimated and what can be learnt from them. This is different from fitting models merely for prediction purposes, such as when predicting finite population totals or in small area estimation, where the structure and interpretation of the model are of secondary importance. Models are also used implicitly for choosing the sampling design and estimators, such as in stratified sampling, or when defining weighting cells for nonresponse adjustments. However, inference is typically based in these cases on the randomization distribution over all possible sample selections, and not on the model, which is known as 'model assisted inference'.

Survey data typically differ from other data sets in five main aspects.

1. The samples are selected at random with known selection probabilities, which allows using the randomization distribution over all possible sample selections as the basis for inference instead of the hypothetical distribution underlying the population model. As discussed below, a combination of the two distributions is in common use.
2. The sample selection probabilities in at least some stages of the sample selection are often unequal; when these probabilities are related to the model outcome variable, the sampling process becomes informative and the model holding for the sample is then different from the target population model.
3. Survey data are almost inevitably subject to various forms of nonresponse, often of considerable magnitude, which again may distort the population model if the response propensity is associated with the outcome of interest (not missing at random nonresponse).
4. The sample data are often clustered due to the use of multi-stage cluster samples. The clusters are 'natural units' (households, individuals in case of longitudinal surveys…), implying that observations within the same cluster are correlated.
5. The data available to the modeler may be masked ("swapped", "contaminated", suppressed") in order to protect the anonymity of the respondents. When this is the case, the modeler's data differ from the correct data.

Many approaches have been proposed in the literature for estimating population models from complex survey data possessing these features, some of which are more familiar than the others. The approaches differ in the conditions underlying their use, the data required for their application, goodness of fit testing, the inference objectives that they accommodate, statistical efficiency, computational demands, and the skills required from analysts fitting the model. This heterogeneity means that there does not exists any single

[1] Danny Pfeffermann, Southampton Statistical Sciences Research Institute, U.K. and Hebrew University of Jerusalem, Israel. E-mail: d.pfeffermann@soton.ac.uk.

approach that can be considered as best in all situations. That being the case, a fundamental question arising is which approach or approaches could or should be used for a given practical application.

The present paper is divided into three parts. In the first part (Section 2) I elaborate on the first four features of complex survey data mentioned above. In the second part (Section 3) I review the various approaches proposed in the literature for dealing with these features, discussing their merits and limitations in light of the properties mentioned above. In the third part (Section 4) I present simulation results which compare the approaches when estimating a linear regression model from a stratified sample in terms of bias, variance, and coverage rates. I conclude with a short discussion of pending issues in Section 5.

2. Why are survey data different from other data?

2.1 The problem of unequal sampling probabilities and nonresponse

Consider a finite population $U = \{1, ..., N\}$ with measurements $\{y_i, \mathrm{x}_i, \mathrm{z}_i\}$ for unit $i = 1, ..., N,$ where y represents an outcome variable of interest, x a vector of covariates and z a vector of design variables used for the sample selection. The design variables may include some or all of the covariates, and in special cases also the outcome variable when known for all the population units, such as in case-control studies. The matrix $Z_U = [\mathrm{z}_1, ..., \mathrm{z}_N]$ is known to the sampler drawing the sample, but not necessarily to the analyst fitting the model. Denote by $s = (\mathrm{I}_1, ..., \mathrm{I}_N)$ the selected sample, where I_i is the sampling indicator taking the value 1 if unit $i \in U$ is drawn to the sample and 0 otherwise. In practice, not all the sampled units necessarily respond, and we denote by R_i the response indicator; $R_i = 1(0)$ if unit $i \in S$ responds (does not respond).

The observed data may be viewed as the outcome of three random processes. The first process generates the vectors $\{y_i, \mathrm{x}_i, \mathrm{z}_i\}$ for the N population units. The second process selects a sample s from U at random by a sampling design, $\Pr(s) = \Pr(s \mid Z_U)$. The third process selects the responding units. This process is obviously not part of the original sampling design and is often the result of 'self selection', although nonresponse could be caused by many other reasons. See Brick and Montaquila (2009) for a recent overview.

When the sample selection probabilities and/or the response probabilities are related to the values of the outcome variable even after conditioning on the model covariates, in the sense that $\Pr(\mathrm{I}_i = 1 \mid y_i, \mathrm{x}_i) \neq \Pr(\mathrm{I}_i = 1 \mid \mathrm{x}_i)$ or $\Pr(R_i = 1 \mid y_i, \mathrm{x}_i, \mathrm{I}_i = 1) \neq \Pr(R_i \mid \mathrm{x}_i, \mathrm{I}_i = 1),$ the model holding for the observed outcomes is different from the population model. In symbols, $f_o(y_i|\mathrm{x}_i) \neq f_p(y_i|\mathrm{x}_i),$ where $f_o(y_i|\mathrm{x}_i)$ represents the model holding for a unit selected to the sample and responding, and $f_p(y_i|\mathrm{x}_i)$ is the *population* model (the model holding for the population values). See Equations (2.1) and (2.2) below.

Example 1. Suppose that the population model is the regression model, $f_p(y_i \mid \mathrm{x}_i) = N(\mathrm{x}_i'\beta, \sigma_\varepsilon^2),$ and that the sample is selected with selection probabilities satisfying $\Pr(I_i = 1 \mid y_i, \mathrm{x}_i) = \exp[\gamma_1 y_i + \gamma_2 y_i^2 + g(\mathrm{x}_i)],$ where γ_1 and $\gamma_2 \leq 0$ are constants and $g(\mathrm{x}_i)$ is some nonstochastic function of the covariates. Simple use of Bayes theorem (see below) shows that the model holding for the sample outcomes is in this case, $f_s(y_i \mid \mathrm{x}_i) = N[(\gamma_1\sigma^2 + \mathrm{x}_i'\beta)/C,$ $\sigma_\varepsilon^2/C],$ where $C = (1 - 2\sigma_\varepsilon^2\gamma_2).$ Thus, although the sample residuals have again a normal distribution, the regression coefficients and the residual variance are different from their values under the population model. In the special case $\gamma_2 = 0,$ the slope coefficients and the residual variance are the same as under the population model, but not the intercept. If $\gamma_1 = 0$ as well, the sample selection probabilities satisfy $\Pr(\mathrm{I}_i = 1 \mid y_i, \mathrm{x}_i) = \Pr(\mathrm{I}_i = 1 \mid \mathrm{x}_i)$ and the two models are now the same.

Following conventional terminology, when $\Pr(\mathrm{I}_i = 1 \mid y_i, \mathrm{x}_i) \neq \Pr(\mathrm{I}_i = 1 \mid \mathrm{x}_i)$ the sampling design is said to be *informative*. When $\Pr(R_i = 1 \mid y_i, \mathrm{x}_i, \mathrm{I}_i = 1) \neq \Pr(R_i \mid \mathrm{x}_i, \mathrm{I}_i = 1),$ the nonresponse is *not missing at random* (NMAR nonresponse). Notice that whereas the sampling probabilities are typically known to the analyst fitting the model, at least for the sampled units, the response probabilities are generally unknown and need to be modelled under NMAR nonresponse. Ignoring an informative sample or NMAR nonresponse and thus assuming implicitly that the model holding for the observed outcomes is the same as the target population model may yield large biases and erroneous inference. The books edited by Kasprzyk, Duncan, Kalton and Singh (1989), Skinner, Holt and Smith (1989) and Chambers and Skinner (2003) contain many discussions and illustrations of the effect of ignoring informative sampling or NMAR nonresponse. See also Pfeffermann (1993, 1996), Pfeffermann and Sverchkov (2009) and Pfeffermann and Sikov (2011) for further discussions and examples, with many other more recent references.

In what follows, I use the abbreviation "*pdf*" to define the probability density function when the outcome is continuous or the probability function when the outcome is discrete. Suppose first that there is no nonresponse. Following Pfeffermann, Krieger and Rinott (1998a), the *marginal sample pdf*, $f_s(y_i \mid \mathrm{x}_i)$ defines the conditional *pdf* of y_i given that unit i is in the sample $(\mathrm{I}_i = 1).$ By Bayes theorem,

$$f_s(y_i \mid \mathrm{x}_i) = f(y_i \mid \mathrm{x}_i, I_i = 1) = \frac{\Pr(I_i = 1 \mid \mathrm{x}_i, y_i) f_p(y_i \mid \mathrm{x}_i)}{\Pr(I_i = 1 \mid \mathrm{x}_i)}, \quad (2.1)$$

where $f_p(y_i \mid \mathrm{x}_i)$ is the corresponding population *pdf*. The probabilities $\Pr(I_i = 1 \mid \mathrm{x}_i, y_i)$ are generally not the same as the sample selection probabilities $\pi_i = \Pr(\mathrm{I}_i = 1)$, which may depend on all the population values Z_U of the design variables. However, the use of the marginal sample *pdf* only requires modelling $\Pr(I_i = 1 \mid \mathrm{x}_i, y_i)$. Typically, $\Pr(I_i = 1 \mid \pi_i, y_i, \mathrm{x}_i) = \pi_i$, in which case $\Pr(I_i = 1 \mid y_i, \mathrm{x}_i) = E_p(\pi_i \mid y_i, \mathrm{x}_i)$, where $E_p(\cdot)$ is the expectation under the population *pdf*.

Remark 1. In practice, the covariates featuring in the population model need not be the same as the covariates featuring in the model of the conditional sample inclusion probabilities, $\Pr(I_i = 1 \mid \mathrm{x}_i, y_i)$. In fact, following the results in Pfeffermann and Landsman (2011), identifiability of the sample model often requires that the two sets of covariates are not identical. However, to simplify the presentation in this paper, I assume for convenience that the covariates contained in the population model and the covariates defining the conditional inclusion probabilities are the same, or alternatively, that x_i defines the union of the two sets of covariates.

It follows from (2.1) that unless $\Pr(I_i = 1 \mid \mathrm{x}_i, y_i) = \Pr(I_i = 1 \mid \mathrm{x}_i) \forall y_i$, the sample *pdf* is different from the population *pdf*, in which case the sampling design is informative and cannot be ignored in the inference process. In particular, it follows from (2.1) that under informative sampling,

$$E_s(y_i \mid \mathrm{x}_i) = E_p\left[\frac{\Pr(I_i = 1 \mid \mathrm{x}_i, y_i) y_i}{\Pr(I_i = 1 \mid \mathrm{x}_i)} \middle| \mathrm{x}_i\right] \neq E_p(y_i \mid \mathrm{x}_i),$$

where $E_s(\cdot)$ is the expectation under the sample *pdf*. Estimating $E_p(y_i \mid \mathrm{x}_i)$ is often the main target of inference, illustrating that ignoring an informative sampling scheme and thus estimating implicitly $E_s(y_i \mid \mathrm{x}_i)$ can bias the inference.

Suppose now the existence of NMAR nonresponse. The marginal sample *pdf* (2.1) can be extended to this case by defining,

$$\begin{aligned} f_o(y_i \mid \mathrm{x}_i) &= f(y_i \mid \mathrm{x}_i, I_i = 1, R_i = 1) \\ &= \frac{\Pr(R_i = 1 \mid y_i, \mathrm{x}_i, I_i = 1) \Pr(I_i = 1 \mid y_i, \mathrm{x}_i) f_p(y_i \mid \mathrm{x}_i)}{\Pr(R_i = 1 \mid \mathrm{x}_i, I_i = 1) \Pr(I_i = 1 \mid \mathrm{x}_i)} \\ &= \frac{\Pr(R_i = 1 \mid y_i, \mathrm{x}_i, I_i = 1) f_s(y_i \mid \mathrm{x}_i)}{\Pr(R_i = 1 \mid \mathrm{x}_i, I_i = 1)}. \end{aligned} \qquad (2.2)$$

Notice from (2.2) that unless $\Pr(R_i = 1 \mid y_i, \mathrm{x}_i, \mathrm{I}_i = 1) = \Pr(R_i \mid \mathrm{x}_i, \mathrm{I}_i = 1) \forall y_i$, the *pdf* holding for the observed outcomes is different from the sample *pdf*. Here again I assume for convenience that the response probabilities depend on the same covariates as in the sample model. See Remark 1 above.

The *pdf*s (2.1) and (2.2) define the marginal distributions of the outcome for a given unit. These definitions generalize very naturally to the joint *pdf* of two or more outcomes associated with different units. More generally, define for every plausible sample $s \subset U$ the sample indicator A_s, such that $A_s = 1$ if s is sampled and $A_s = 0$ otherwise, and assume for convenience full response. Denote the data associated with s by $(\mathrm{y}_s, \mathrm{x}_s)$. The joint sample *pdf* of $\mathrm{y}_s \mid \mathrm{x}_s$ is then,

$$\begin{aligned} f_s(\mathrm{y}_s \mid \mathrm{x}_s) &= f(\mathrm{y}_s \mid \mathrm{x}_s, A_s = 1) \\ &= \frac{\Pr(A_s = 1 \mid \mathrm{y}_s, \mathrm{x}_s) f_p(\mathrm{y}_s \mid \mathrm{x}_s)}{\Pr(A_s = 1 \mid \mathrm{x}_s)}. \end{aligned} \qquad (2.3)$$

The *pdf* $f_p(\mathrm{y}_s \mid \mathrm{x}_s)$ can be general, allowing in particular for correlated measurements, but modelling the probability $\Pr(A_s = 1 \mid \mathrm{y}_s, \mathrm{x}_s)$ is practically only feasible if the sample can be decomposed into exclusive and exhaustive subsets s_k such that $\Pr(A_s = 1 \mid \mathrm{y}_s, \mathrm{x}_s) \propto \prod_k \Pr(A_{s_k} = 1 \mid \mathrm{y}_{s_k}, \mathrm{x}_{s_k})$ and $\Pr(A_{s_k} = 1 \mid \mathrm{y}_{s_k}, \mathrm{x}_{s_k})$ satisfies the same model for all the subsets (see Example 2). In particular, if the population outcomes are independent given the covariates under the population model and $\Pr(A_s = 1 \mid \mathrm{y}_s, \mathrm{x}_s) \propto \prod_{i \in s} \Pr(\mathrm{I}_i = 1 \mid y_i, \mathrm{x}_i)$, (2.3) takes the form

$$\begin{aligned} f_s(\mathrm{y}_s \mid \mathrm{x}_s) &= \prod\nolimits_{i \in s} \frac{\Pr(\mathrm{I}_i = 1 \mid y_i, \mathrm{x}_i) f_p(y_i \mid \mathrm{x}_i)}{\Pr(\mathrm{I}_i = 1 \mid \mathrm{x}_i)} \\ &= \prod\nolimits_{i \in s} f_s(y_i \mid \mathrm{x}_i), \end{aligned} \qquad (2.4)$$

so that the sample outcomes are likewise independent.

Example 2. Consider the case of a clustered population $U = \bigcup_l U_l$, with independent measurements between clusters, such that $f_p(\mathrm{y}_U \mid \mathrm{x}_U) = \prod_l f_p(\mathrm{y}_{U_l} \mid \mathrm{x}_{U_l})$, where $(\mathrm{y}_U, \mathrm{x}_U)$ defines all the population values and $(\mathrm{y}_{U_l}, \mathrm{x}_{U_l})$ the values in cluster l. Let s define the set of sampled clusters, assumed to be drawn independently with probabilities $\Pr(l \in s \mid \mathrm{y}_{U_l}, \mathrm{x}_{U_l}) = r(\mathrm{y}_{U_l}, \mathrm{x}_{U_l})$ for some function $r(\cdot)$, and suppose also that all the units in the sampled clusters are observed (single-stage cluster sampling). Then, $\Pr(A_s = 1 \mid \mathrm{y}_U, \mathrm{x}_U) = \prod_{k \in s} r(\mathrm{y}_{U_k}, \mathrm{x}_{U_k}) \times \prod_{j \notin s}[1 - r(\mathrm{y}_{U_j}, \mathrm{x}_{U_j})]$. Since for $k \in s, (\mathrm{y}_{U_k}, \mathrm{x}_{U_k}) = (\mathrm{y}_{s_k}, \mathrm{x}_{s_k})$, it follows that $\Pr(A_s = 1 \mid \mathrm{y}_s, \mathrm{x}_s) = \prod_{k \in s} r(\mathrm{y}_{s_k}, \mathrm{x}_{s_k}) \times G$, where for given covariates x_{U_j}, $j \notin s_1$, G is a constant satisfying, $G = \prod_{j \notin s} \int [1 - r(\mathrm{y}_{U_j}, \mathrm{x}_{U_j})] f_p(\mathrm{y}_{U_j} \mid \mathrm{x}_{U_j}) d\mathrm{y}_{U_j}$. The case of a non-clustered population with independent measurements and Poisson sampling of individual units is a special case where each cluster consists of a single element, giving rise to (2.4).

Remark 2. The examples considered so far assume independent sampling, which preserves the independence of the outcomes after sampling, but this assumption can usually be relaxed following a result proved and illustrated in Pfeffermann *et al.* (1998a). By this result, under some general regularity conditions and for many commonly used sampling schemes for selection with unequal probabilities, if the population measurements are independent, the sample measurements are *asymptotically independent* under the sample distribution. The asymptotic framework requires that the population size increases but the sample size is held fixed. As illustrated in section 2.3, the assumption of independent population measurements is often also not restrictive.

So far, we suppressed for convenience from the notation the parameters underlying the population *pdf* and the sampling process. Consider, for example, the sample *pdf* (2.3). With added parameter notation, it can be written as

$$f_s(\mathrm{y}_s \,|\, \mathrm{x}_s;\, \theta,\, \gamma) = \frac{\Pr(A_s = 1 \,|\, \mathrm{y}_s,\, \mathrm{x}_s;\, \gamma)\, f_p(\mathrm{y}_s \,|\mathrm{x}_s;\, \theta)}{\Pr(A_s = 1 \,|\mathrm{x}_s;\, \theta,\, \gamma)}. \quad (2.5)$$

Thus, the conditional population and sample *pdf*s are different, unless

$$\Pr(A_s = 1 \,|\, \mathrm{y}_s,\, \mathrm{x}_s;\, \gamma) = \Pr(A_s = 1 \,|\, \mathrm{x}_s;\, \theta,\, \gamma)\ \forall\, \mathrm{y}_s. \quad (2.6)$$

When (2.6) holds, inference on the target parameter θ can be implemented by fitting the population model to the sample data, ignoring the sample selection. Note that this conclusion refers to the selected sample defined by the event $A_s = 1$.

The condition (2.6) is a strong condition. In a fundamental article on missing values, Rubin (1976) establishes conditions under which the sampling process can be ignored for likelihood, Bayesian or sampling theory (repeated sampling from a model) inference, that is, conditions under which the population model defined by $f_p(\mathrm{y}_s \mid \mathrm{x}_s;\, \theta)$ can be fitted to the observed data, depending on the inference method used. Little (1982) extends Rubin's results by distinguishing between the sample selection and the response process. Another important distinction is that Little conditions on the population values Z_U of the design variables used for the sample selection. Inference on the target population model $f_p(\mathrm{y}_s \mid \mathrm{x}_s;\, \theta)$ requires therefore integrating the conditional *pdf* of $\mathrm{y}_s \mid Z_U,\, \mathrm{x}_s$ over the distribution of $Z_U \mid \mathrm{x}_s$ (see Section 3). Sugden and Smith (1984) establish conditions under which a sampling process that depends on design variables Z is ignorable, given partial information on the design. Let $d_s = D_s(z_U)$ contain all the available design information for a sample s such as strata membership (may only be known for the sampled units), sample selection probabilities *etc.* Using previous notation, a key condition for ignorability of the sampling process given the available design information is that $A_s \perp Z_U \mid d_s$, with "$\perp$" meaning independence, implying $\Pr(A_s = 1 \mid Z_U = z_U) = \Pr(A_s = 1 \mid d_s)$ for any z_U for which $D_s(z_U) = d_s$.

For large scale multi-stage sample surveys with possibly many design variables, it is generally difficult and often impractical to check directly the conditions that permit ignoring the sample selection or nonresponse given the available design information. On the other hand, even when the sample *pdf* is different from the population *pdf*, it does not necessarily imply that inference that ignores the sampling process is wrong. As a simple illustration, consider the special case of Example 1 where $\gamma_2 = 0$. In this case the sample *pdf* is normal with the same slope coefficients and residual variance as under the population *pdf*. Thus, for inference about the slope coefficients one can ignore the sampling process. A similar result holds for logistic models when the sample selection depends on y but not on x. See Pfeffermann *et al.* (1998a) for derivation of this result. Pfeffermann and Sverchkov (2009) review several test statistics proposed in the literature for assessing whether ignoring the sample selection is justified for the intended inference.

2.2 The use of the randomization distribution for inference

A unique feature of sample surveys is that the sample is selected at random by use of a sampling design $[\{s, \Pr(s)\},\ s \in S]$. The sampling design induces a (discrete) *randomization distribution* for any statistic T_{ys}, which is the conditional distribution over all possible sample selections, given the finite population values. Thus, the statistic T_{ys} takes the value t_{ys} with probability $\Pr(s),\ s \in S$. Classical survey sampling inference is based solely on this distribution. For example, the familiar Horvitz-Thompson (HT) estimator T_{ys}^{HT}, which takes the value $t_{ys}^{\mathrm{HT}} = \sum_{i \in s}(y_i / \pi_i)$ if sample s is drawn, is randomization-unbiased for the finite population total $\mathrm{TOT}_y = \sum_{j=1}^{N} y_j$, since $\sum_{s \in S} \Pr(s)\, t_{ys}^{\mathrm{HT}} = T_y$. Its variance is, $\mathrm{Var}(T_{ys}^{\mathrm{HT}}) = \sum_{s \in S} \Pr(s) (t_{ys}^{\mathrm{HT}} - T_y)^2$. Notice that in the case of nonresponse, the use of the randomization distribution requires knowledge of the response probabilities, which in practice can only be estimated. The HT estimator takes in this case the form, $T_{ys}^{\mathrm{HT}} = \sum_{i \in R} y_i / [\pi_i \times \hat{\Pr}(R_i = 1 \,|\, I_i = 1)]$, where R defines the subsample of respondents. See Fuller (2002) for further discussion.

The randomization distribution conditions on the realized population values. Consequently, it can be used for descriptive inference on known functions of the finite population values, but not for analytic inference on a hypothesized model giving rise to these values. For this, one may consider the joint distribution over all possible sample outcomes for given population values (the *randomization r*-distribution) and all possible realizations of the finite population measurements (the *model p*-distribution). See Binder and Roberts (2009) and the references therein. The combined $r-p$ distribution offers an alternative framework of inference to the use of the *pdf*s $f_s(y \mid \mathrm{x})$ or $f_o(y \mid \mathrm{x})$ defined before.

Example 3: Suppose that the population model is $y_i \sim \mathrm{Mult}[\{p_k\}, K]$, such that $\Pr_p(y_i = k) = p_k, k = 1, ..., K$; $\sum_{k=1}^{K} p_k = 1.$ Let $\Pr(i \in s \mid y_i = k) = \pi_k$. Then, by (2.1), $\Pr_s(y_i = k) = \Pr(y_i = k \mid i \in s) = \pi_k p_k / \sum_{j=1}^{K} \pi_j p_j = p_k^*$, or, $y_i \mid i \in s \sim \mathrm{Mult}(\{p_k^*\}, K)$. Assuming independence of the observed outcomes and known selection probabilities, the maximum likelihood estimator (*mle*) of p_k based on the sample distribution is $\tilde{p}_k = (n_k / \pi_k) / \sum_{j=1}^{K} (n_j / \pi_j)$, where n_k is the number of sampled units with outcome $y_i = k$. The use of the $r - p$ distribution suggests estimating p_k by the HT estimator $\hat{p}_k = (1/N) \sum_{i \mid y_i = k} (1/\pi_k) = (n_k / \pi_k)/N$. The estimator $\hat{p}_k$ is randomization-unbiased for $\hat{P}_k = N_k / N$, where N_k is the number of population units with outcome $y_j = k$, and $\hat{P}_k$ is *p*-unbiased for p_k, such that $\hat{p}_k$ is $r - p$-unbiased for p_k.

The obvious difference between the $r - p$ distribution and the sample distribution, $f_s(y \mid \mathrm{x})$, is that the latter conditions on the observed sample of units (and hence the observed values of the covariates or the selected clusters in a cluster sample), whereas the $r - p$ distribution accounts for all possible sample selections. Consequently, the use of the latter distribution does not lend itself in general to conditional inference. The use of the *pdf*s $f_s(y \mid \mathrm{x})$ or $f_o(y \mid \mathrm{x})$ requires modelling $\Pr(I_i = 1 \mid \mathrm{x}_i, y_i)$ (Equation 2.1) and $\Pr(R_i = 1 \mid y_i, \mathrm{x}_i, I_i = 1)$ in case of nonresponse (Equation 2.2), but it permits the computation (estimation) of the conditional *pdf* of the observed outcomes given the covariates, and hence the use of classical inference tools.

2.3 Data obtained from a cluster sample

Another special feature of survey data mentioned in the introduction is *clustering*, due to the use of multi-stage cluster samples. The clusters are ‘natural groups’ such as households, residence blocks, schools, or even individuals in the case of longitudinal surveys. Consequently, the outcomes pertaining to the same cluster are generally correlated, known as the *intraclass correlation*. It is important to emphasize that the clusters represent an existing population grouping, such that an intraclass correlation exists also under the population model.

Pfeffermann and Smith (1985) review several classes of plausible regression models for clustered populations, and discuss how they can be estimated from the sample. A population model in common use is the random intercept model,

$$y_{ij} = \mathrm{x}'_{ij} \beta + u_i + \varepsilon_{ij};\ i = 1, ..., M,\ j = 1, ..., N_i;$$
$$u_i \overset{\text{indep.}}{\sim} N(0, \sigma_u^2);\ \varepsilon_{ij} \overset{\text{indep.}}{\sim} N(0, \sigma_\varepsilon^2), \tag{2.7}$$

where M defines the number of clusters and N_i the number of units in cluster *i*. The model assumes also $E(u_i \varepsilon_{ij}) = 0, \forall i, j$. Notice that under this model $\mathrm{Var}(y_{ij}) = \sigma_u^2 + \sigma_\varepsilon^2$, $E(y_{ij} y_{il}) = \sigma_u^2$ for $j \neq l$ and $E(y_{ij} y_{kl}) = 0$ for $i \neq k$, implying

$$\begin{aligned} \mathrm{Corr}(y_{ij}, y_{il}) &= \sigma_u^2 / (\sigma_u^2 + \sigma_\varepsilon^2) \text{ for } j \neq l; \\ \mathrm{Corr}(y_{ij}, y_{kl}) &= 0 \text{ for } i \neq k. \end{aligned} \tag{2.8}$$

Scott and Holt (1982) show that estimating β in (2.7) by ordinary least squares (OLS) usually results in a small loss of efficiency, compared to the use of the optimal generalized least squares (GLS) estimator. However, ignoring the intracluster correlation when estimating the variance of the OLS estimator may result in considerable variance underestimation and hence wrong size and excessive powers of test statistics and too short confidence intervals.

The results in Scott and Holt (1982) and Pfeffermann and Smith (1985) assume noninformative sampling and full response. When this is not the case, the model holding for the sample data is different from the corresponding population model, although the clustered nature of the model is preserved as we now show. Consider the following two-level population model:

$$\begin{aligned} &\text{Level 1: } u_i \mid \mathrm{t}_i \sim \varphi_p(u_i \mid \mathrm{t}_i; \theta_1),\ i = 1,...,M \\ &\text{Level 2: } Y_{ij} \mid (u_i, \mathrm{x}_{ij}) \sim f_p(y_{ij} \mid \mathrm{x}_{ij}, u_i; \theta_2),\ j = 1,...,N_i, \end{aligned} \tag{2.9}$$

where φ_p and f_p denote the first and second-level *pdf*s with known covariates t_i and x_{ij}, governed by the hyperparameters θ_1 and θ_2 respectively. The model (2.7) is a special case of (2.9) by which φ_p and f_p are normal *pdf*s with $\mathrm{t}_i = 0$ (no covariates), $\theta_1 = \sigma_u^2$ and $\theta_2 = (\beta, \sigma_\varepsilon^2)$. Suppose that the sample is drawn by the following two-stage sampling process. In the first stage a sample s_1 of $m < M$ first-level units (clusters; say, schools) is selected with probabilities $\pi_i = \Pr(i \in s_1)$ that may be correlated with the random effects u_i after conditioning on the covariates t_i. In the second stage a sub-sample s_{2i} of $n_i < N_i$ second-level units (ultimate sampling units; say, pupils) is sampled from each selected first-level unit i with probabilities $\pi_{j|i} = \Pr(j \in s_{2i} \mid i \in s_1)$ that may be correlated with the outcomes y_{ij} after conditioning on the covariates x_{ij}. Denote by I_i and $\mathrm{I}_{j|i}$ the first and second-stage sampling indicators. By (2.1), the two-level sample model holding for the observed data, corresponding to the population model (2.9) is,

$$\begin{aligned} &\text{Level 1:} \\ &f_{s_1}(u_i \mid \mathrm{t}_i; \theta_1, \gamma_1) \\ &\quad = \frac{\Pr(\mathrm{I}_i = 1 \mid u_i, \mathrm{t}_i; \gamma_1) \varphi_p(u_i \mid \mathrm{t}_i; \theta_1)}{\Pr(\mathrm{I}_i = 1 \mid \mathrm{t}_i; \theta_1, \gamma_1)} \\ &\text{Level 2:} \\ &f_{s_{2i}}(y_{ij} \mid \mathrm{x}_{ij}, u_i; \theta_2, \gamma_2) \\ &\quad = \frac{\Pr(\mathrm{I}_{j|i} = 1 \mid y_{ij}, \mathrm{x}_{ij}; \gamma_2) f_p(y_{ij} \mid \mathrm{x}_{ij}, u_i; \theta_2)}{\Pr(\mathrm{I}_{j|i} = 1 \mid u_i, \mathrm{x}_{ij}; \theta_2, \gamma_2)}, \end{aligned} \tag{2.10}$$

where I assume $\Pr(\mathrm{I}_{j|i} = 1 \mid y_{ij}, u_i, \mathrm{x}_{ij}; \gamma_2) = \Pr(\mathrm{I}_{j|i} = 1 \mid y_{ij}, \mathrm{x}_{ij}; \gamma_2)$.

Remark 3. By the independence result in Remark 2, if $y_{ij} \mid u_i$ are independent under the population model, they are asymptotically independent under the sample model. Similarly, if the random effects u_i are independent under the population model, they are asymptotically independent under the sample model. Thus, the sample model (2.10) is a genuine two-level model, although with different distributions and possibly more parameters. Evidently, the models (2.9) and (2.10) are different, unless $\Pr(\mathrm{I}_{j|i} = 1 \mid y_{ij}, \mathrm{x}_{ij}) = \Pr(\mathrm{I}_{j|i} = 1 \mid u_i, \mathrm{x}_{ij})$ and $\Pr(\mathrm{I}_i = 1 \mid u_i, \mathrm{t}_i) = \Pr(\mathrm{I}_i = 1 \mid \mathrm{t}_i)$.

So far I assumed implicitly full response. Suppose, for example, that in sampled cluster (first level unit) i only a sub-sample $r_{2i} \subset s_{2i}$ respond, and denote by $R_{j|i}$ the response indicator. The second-level model for the observed outcomes is now,

Level 1:

$$f_{o2i}(y_{ij} \mid \mathrm{x}_{ij}, u_i; \theta_2, \gamma_2, \gamma_2^*)$$

$$= f(y_{ij} \mid \mathrm{x}_{ij}, u_i, I_{j|i} = 1, R_{j|i} = 1)$$

$$= \frac{\Pr(R_{j|i} = 1 \big| y_{ij}, \mathrm{x}_{ij}, I_{j|i} = 1; \gamma_2^*) f_{s_{2i}}(y_{ij} | \mathrm{x}_{ij}, u_i; \theta_2, \gamma_2)}{\Pr(R_{j|i} = 1 \big| \mathrm{x}_{ij}, u_i, I_{j|i} = 1; \theta_2, \gamma_2, \gamma_2^*)}. \quad (2.11)$$

The *pdf* (2.11) coupled with the level 1 *pdf* in (2.10) defines the model holding for the observed data in the case of informative cluster sampling and NMAR nonresponse.

3. How can we estimate population models from complex survey data?

In this section I review the main approaches proposed in the literature to deal with the special features of complex survey data discussed in Section 2, and propose some modifications. In order to simplify the discussion, I consider the following set up used for the simulation study in Section 4.

3.1 Population model and sampling design

Consider a stratified population $U = U_1 \cup ... \cup U_H$ of size N. Specifically, define for every unit $j \in U$ a random vector stratification indicator $\mathrm{z}_j = (z_{1j}, ..., z_{Hj})'$ such that $\Pr(z_{hj} = 1) = p_h, \sum_{h=1}^{H} p_h = 1$ and $j \in U_h$ if $z_{hj} = 1$. The stratification is carried out independently between the units. Values of an outcome variable Y are generated as $y_j = \beta_0 + \beta_1 x_j + \alpha_0 \zeta_j + \alpha_1 \zeta_j x_j + \varepsilon_j$; $\varepsilon_j \sim N(0, \sigma^2)$, where the x_j's are fixed scalar covariates, $(\beta_0, \beta_1, \alpha_0, \alpha_1)$ are fixed coefficients and

$$\zeta_j = \frac{1}{H} \sum\nolimits_{h=1}^{H} \frac{z_{hj}}{p_h} - 1.$$

Notice that ζ_j is a random variable with mean zero and variance

$$V_\zeta = \left(\frac{1}{H^2} \sum\nolimits_{h=1}^{H} \frac{1}{p_h} \right) - 1,$$

implying that for given covariates x_j, x_k,

$$E_p(y_j \mid x_j)$$

$$= \beta_0 + \beta_1 x_j, \ \mathrm{Var}_p(y_j \mid x_j)$$

$$= (\alpha_0 + \alpha_1 x_j)^2 V_\zeta + \sigma^2, \ \mathrm{Cov}_p(y_j, y_k \mid x_j, x_k)$$

$$= 0, \ j \neq k. \quad (3.1)$$

However, for unit $j \in U_h$,

$$y_j \mid x_j, \mathrm{z}_{hj} = 1 \sim N[(\beta_0 + \alpha_0 \zeta_h)$$

$$+ (\beta_1 + \alpha_1 \zeta_h) x_j, \sigma^2]; \ \zeta_h = [(1/Hp_h) - 1]. \ (3.2)$$

Thus, the regression model in each stratum is the classical linear model with constant variance, but the intercepts and slopes change across the strata.

The model defined by (3.1) and (3.2) is a realistic random coefficients regression model, which I think mimics many populations encountered in practice.

We used systematic probability proportional to size (PPS) sampling within the strata for drawing the samples with the size variable defined as $z_j^* = \max\{\min[(|q_j|)^{1.5}, 9], 1\}$; $q_j \sim N(1 + x_j, 1)$. There is nothing novel about the choice of this size variable except that it allows for a clear distinction between the variance of the various estimators. This size z_j^* does not depend on the outcome y_j, and hence the sampling process within each stratum is non-informative. However for disproportionate allocation of the sample between the strata, the sampling scheme is informative because of the different models operating in different strata, such that the observed outcomes carry information on the strata membership and $\Pr(j \in s \mid y_j, x_j) \neq \Pr(j \in s \mid x_j)$. We focus on the estimation of the regression coefficients (β_0, β_1) in (3.1) as the target of inference and assume that the available sample information consists of the observed outcomes and covariates, the strata membership vectors z_{hj} and the strata sizes, $\{N_h\}$.

3.2 Including the design variables among the covariates

As implied by (2.3), the population model (*pdf*), $f_p(\mathrm{y}_s \mid \mathrm{x}_s)$ and the sample model $f_s(\mathrm{y}_s \mid \mathrm{x}_s)$ are the same if $\Pr(A_s = 1 \mid \mathrm{y}_s, \mathrm{x}_s) = \Pr(A_s = 1 \mid \mathrm{x}_s) \forall \mathrm{y}_s$. By (2.2), the response process is ignorable if $\Pr(R_i = 1 \big| y_i, \mathrm{x}_i, I_i = 1) = \Pr(R_i = 1 \big| \mathrm{x}_i, I_i = 1) \forall y_i$. Thus, a possible way to account for the sampling and response effects is to add to the model covariates all the variables and interactions determining the sample and response probabilities and then integrate them out in order to estimate the model of interest. Denote these variables by $J = Z \cup L$ with population values J_U, where L defines the variables explaining the response probabilities. Assuming $f_p(\mathrm{y}_s | \mathrm{x}_U, j_U) = f_p(\mathrm{y}_s | \mathrm{x}_s, j_U)$, the use of this approach requires to fit first the model

$$f_p(\mathrm{y}_s \,|\, \mathrm{x}_s, J_U = j_U) = \int f_p(\mathrm{y}_s, \mathrm{y}_{\bar{s}} \,|\, \mathrm{x}_U, j_U)\, d\,\mathrm{y}_{\bar{s}}, \quad (3.3)$$

and then integrate,

$$f_p(\mathrm{y}_s \,|\, \mathrm{x}_s) = \int f_p(\mathrm{y}_s \,|\, \mathrm{x}_s, j_U)\, f_p(j_U \,|\, \mathrm{x}_s)\, d j_U. \quad (3.4)$$

Variants of this approach can be found in DeMets and Halperin (1977), Holt, Smith and Winter (1980), Nathan and Holt (1980), Jewell (1985), Skinner (1994), Chambers and Skinner (2003, Chapter 2) and Gelman (2007).

The use of the approach is appealing, and it has the advantage of allowing classical model based inference procedures once the variables $J_U = Z_U \cup L_U$ are included in the model, but it is often limited in practice for the following reasons:

1. It requires knowledge of the population values of all the variables determining the sample selection and response, and this information is usually unknown to the analyst fitting the model because of confidentiality restrictions or other reasons. Even if known, including in the model all the geographic and operational variables used for the sampling design and the variables explaining the response may be formidable.
2. In practice there may be many covariates and many design variables, and modelling the relationship between the design variables and the covariates in order to integrate out the effect of the former variables can be complicated and may no longer reproduce the original target model.

Feder (2011) proposes the following simple solution to this problem. Suppose first that the design variables and the covariates are known for every element in the population. The proposed solution consists of imputing the missing population outcomes using the model $f_p(\mathrm{y}_s \,|\, \mathrm{x}_s, J_U = j_U)$ fitted to the sample data, and then fitting the population model $f_p(y_j \,|\, \mathrm{x}_j)$ using all the population values, with the missing outcomes replaced by their imputed values. When the design variables and the covariates are unknown for the non-sampled units, they need to be imputed as well. The imputation may be carried out by sampling with replacement $(N-n)$ values (x_i, z_i) from the sample values with probabilities $\vec{p}_i = (w_i - 1) / \sum_{k=1}^{n} (w_k - 1)$ on each draw, where the w_i 's are the sampling weights. See Pfeffermann and Sikov (2011) for justification of this procedure under the sample model and an extension for the case of NMAR nonresponse.

3. The approach is not operational when the inclusion in the sample depends also on the outcome values, that is, $Z_U = \{Y_U, Z_U^*\}$ and $\Pr(A_s = 1 \,|\, Y_U, X_U, Z_U^*) \neq \Pr(A_s = 1 \,|\, X_U, Z_U^*)$. A classical example is *case-control studies* (Scott and Wild 2009), but a similar problem arises when the nonresponse is NMAR.

Remark 4. Including the design variables and the variables explaining the response in the model does not necessarily require integrating them out even if they are not part of the covariates of interest, as the following example shows.

Example 4: Suppose that a sample of size n is selected with probabilities defined by the population values of design variables Z and that all the sampled units respond. Let the population distribution of Y, X, Z be multivariate normal. The data available to the analyst consist of the sample values $[\mathrm{y}_s, \mathrm{x}_s]$ and the population values Z_U. Using properties of the multivariate normal distribution, $E_p(\mathrm{y} | x) = \beta_0 + \beta_{yx} x$ for some coefficients (β_0, β_{yx}), but the OLS estimate of β_{yx} is biased because the sampling probabilities depend on Z, which is correlated with Y. The *mle* of β_{yx} for the case of a trivariate normal distribution is (DeMets and Halperin 1977),

$$\hat{\beta}_{yx} = \left\{ s_{xy} + \frac{s_{yz} s_{xz}}{s_{zz}} \left(\frac{\hat{\sigma}_z^2}{s_{zz}} - 1 \right) \right\} \Big/ \left\{ s_{xx} + \frac{s_{xz}^2}{s_{zz}} \left(\frac{\hat{\sigma}_z^2}{s_{zz}} - 1 \right) \right\}, \quad (3.5)$$

where $s_{uv} = n^{-1} \sum_{i=1}^{n} (u_i - \bar{u}_s)(v_i - \bar{v}_s)$ and $\hat{\sigma}_z^2 = N^{-1} \sum_{i=1}^{N} (z_i - \bar{z}_U)^2$, with $\bar{u}_s, \bar{v}_s$ and $\bar{z}_U$ defining the corresponding sample and population means. Thus, the population values of Z feature in this case in the optimal estimator of the target parameter β_{yx}. Holt *et al.* (1980) extend this result to the case where Y, X, Z are vector variables. Nathan and Holt (1980) establish conditions under which $\hat{\beta}_{yx}$ is consistent without the multivariate normality assumptions. Pfeffermann and Holmes (1985) study the robustness of the estimator to model misspecification.

3.3 Using the sampling weights as surrogate for the design variables

For situations where there are too many design variables determining the sample selection to include them all in the model, or when some or all of these variables are unknown to the analyst, it is often advocated to include in the model the sampling weights as surrogate of the design variables. Examples of the use of this approach can be found in DuMouchel and Duncan (1983), Särndal and Wright

(1984), Rubin (1985), Chambers, Dorfman and Wang (1998) and Wu and Fuller (2006).

Rubin (1985) defines the vector $a = (a_1, ..., a_N)' = a(Z_U)$ to be an adequate summary of Z_U if $\Pr(A_s = 1 \mid Z_U) = \Pr(A_s = 1 \mid a)$. The author shows that the vector $\pi_U = (\pi_1, ..., \pi_N)$ of the sample inclusion probabilities is the coarsest possible adequate summary of Z_U, though it may be too coarse. It follows therefore that for sampling designs such that $\Pr(A_s = 1 \mid Y_U, Z_U) = \Pr(A_s = 1 \mid Z_U)$, if π_U is an adequate summary, the sample selection can be ignored for inference on the parameters of $f_p(y_s \mid \mathrm{x}_s, \pi_U)$. In order to estimate the target model $f_p(y \mid \mathrm{x})$ in this case, one can follow the same steps as in Section (3.2) with π_U taking the role of Z_U.

The use of this approach reduces the dimension of the added covariates but it requires knowledge of the sample inclusion probabilities (or the sampling weights) for all the population units, which may not be available in the case of a secondary analysis. The case of nonresponse is particularly problematic since the response probabilities are generally unknown and need to be estimated. Another major problem with this approach is that for general sampling designs, the vector π_U may not be an adequate summary of Z. Sugden and Smith (1984) and Smith (1988) establish necessary design information required for sampling ignorability.

Remark 5. Even though the vector π_U is not always an adequate summary of Z_U, for sampling designs such that $\Pr(I_i = 1 \mid y_i, \mathrm{x}_i, \pi_i) = \pi_i$, $f_s(y_i \mid \mathrm{x}_i, \pi_i) = f_p(y_i \mid \mathrm{x}_i, \pi_i)$, so that the marginal population and sample *pdf*s for a given sampled unit are nonetheless the same when adding π_i to the covariates (see Skinner 1994).

Remark 6. In the empirical set up described in Section 3.1 there is a one to one correspondence between the design variables (z'_j, z^*_j) and the sampling weights (w_h, w_j).

3.4 Methods based on probability weighting

So far we considered methods requiring knowledge of the variables J determining the sample selection and response probabilities, or at least an adequate summary of them. The methods considered below only require knowledge of the sampling weights for the responding sampled units. As such, they are restricted to situations of full response, or to cases where the response probabilities can be estimated sufficiently accurately, in which case the sampling weight for a responding unit is the inverse of the product of the unit's selection probability and its estimated response probability. Probability weighting (PW) is discussed in numerous articles; see the recent discussion in Pfeffermann and Sverchkov (2009) and the references therein. As before, we focus here on estimation of population models.

To introduce the idea, consider the case of a *census* with full response. Assuming independent outcomes, the model parameters, θ, are typically estimated in this case by solving *census* estimating equations of the form,

$$\sum\nolimits_{j=1}^{N} u(y_j, \mathrm{x}_j; \theta) = 0. \qquad (3.6)$$

In the case of *mle*, $u(y_j, \mathrm{x}_j; \theta) = (\partial / \partial\theta) \log f_p(\mathrm{y}_j \mid \mathrm{x}_j; \theta)$, the j^{th} score. In practice, data are available for only a sample $s \subset U$ and the equations (3.6) are replaced by their randomization unbiased Horvitz-Thompson estimator,

$$\sum\nolimits_{i\in s} w_i\, u(y_i, \mathrm{x}_i; \theta) = 0, \qquad (3.7)$$

where the w_i 's are the sampling weights.

Remark 7. When the census estimating equations (3.6) are the likelihood equations, the estimators obtained by solving (3.7) are known in the sampling literature as 'pseudo *mle*' (*pmle*). See Binder (1983), Skinner *et al.* (1989), Pfeffermann (1993, 1996) and Godambe and Thompson (2009) for discussion with many examples. This approach is implemented in many software packages such as SAS, STATA, SUDAAN, *etc.*

Example 5. In the case of the standard linear regression model, the *pmle* or PW estimator of the vector coefficient β solves the equations $\sum_{i\in s} w_i (y_i - \mathrm{x}'_i \hat{\beta}_{\text{pw}}) \mathrm{x}_i = 0$;

$$\hat{\beta}_{\text{pw}} = \left[\sum\nolimits_{i\in s} w_i\, \mathrm{x}_i\, \mathrm{x}'_i\right]^{-1} \sum\nolimits_{i\in s} w_i\, \mathrm{x}_i\, y_i. \qquad (3.8)$$

The PW estimator of the residual variance is $\hat{\sigma}^2_{\text{pw}} = \sum_{i\in s} w_i (y_i - \mathrm{x}'_i \hat{\beta}_{\text{pw}})^2 / (\sum_{i\in s} w_i - k)$, where $k = \dim(\beta)$.

For logistic regression, the pseudo likelihood equations (with no explicit solution) are,

$$\begin{aligned} \sum\nolimits_{i\in s} w_i [y_i - \tilde{p}_i(\mathrm{x}_i)]\, \mathrm{x}_i &= 0;\ \tilde{p}_i(\mathrm{x}_i) \\ &= \Pr_p(y_i = 1 \mid \mathrm{x}_i) \\ &= \exp(\mathrm{x}'_i \beta) / [1 + \exp(\mathrm{x}'_i \beta)]. \quad (3.9) \end{aligned}$$

Example 6. Let $u(y_j; \theta) = [\Delta(\theta - y_j) - F_p(\theta)]$ where $F_p(\theta)$ is the cumulative population distribution at θ and $\Delta(a) = 1(0)$ when $a \geq 0\ (a < 0)$. The PW estimator of $F_p(\theta)$ is $\hat{F}_{p,\,\text{pw}}(\theta) = \sum_{i\in s} w_i \Delta(\theta - y_i) / \sum_{i\in} w_i$, the familiar Hàjek (1971) estimator.

The notable property of PW estimators is that they are generally $r - p$ consistent. (See Section 2.2 for definition of the $r - p$ distribution). This can be seen by decomposing $(\hat{\theta}_{\text{pw}} - \theta) = (\hat{\theta}_{\text{pw}} - \hat{\theta}_{\text{cen}}) + (\hat{\theta}_{\text{cen}} - \theta)$, where $\hat{\theta}_{\text{cen}}$ is the (hypothetical) solution of the census equations (3.6). Under general conditions, $(\hat{\theta}_{\text{pw}} - \hat{\theta}_{\text{cen}}) = O_p(n^{-0.5})$ and $(\hat{\theta}_{\text{cen}} - \theta) = O_p(N^{-0.5})$, thus establishing the $r - p$ consistency of $\hat{\theta}_{\text{cen}}$ under these conditions. The $r - p$ variance of $\hat{\theta}_{\text{pw}}$ can be decomposed as,

$$\mathrm{Var}_{r-p}(\hat{\theta}_{\mathrm{pw}}) = E_p[\mathrm{Var}_r(\hat{\theta}_{\mathrm{pw}})] + \mathrm{Var}_p[E_r(\hat{\theta}_{\mathrm{pw}})]. \quad (3.10)$$

For single stage sampling, if n is much smaller than N as is usually the case, the second term on the right hand side of (3.10) is negligible compared to the first term, and $\mathrm{Var}_{r-p}(\hat{\theta}_{\mathrm{pw}})$ can be estimated by the randomization variance estimator $\hat{\mathrm{Var}}_r(\hat{\theta}_{\mathrm{pw}})$. This result does not necessarily hold for cluster sampling since in this case $\mathrm{Var}_r(\hat{\theta}_{\mathrm{pw}})$ is typically of order $O(1/m)$ where m is the number of sampled clusters, and under a suitable model $\mathrm{Var}_p[E_r(\hat{\theta}_{\mathrm{pw}})]$ is $O(1/M)$ where M is the number of population clusters. For $\hat{\mathrm{Var}}_r(\hat{\theta}_{\mathrm{pw}})$ to be an adequate estimator of $\mathrm{Var}_{r-p}(\hat{\theta}_{\mathrm{pw}})$ in this case, m must be much smaller than M.

Remark 8. The consistency of PW estimators under correct population model specification may also be established under the sample distribution (Equation 2.1). Consider the estimator $\hat{\beta}_{\mathrm{pw}}$ in (3.8) and write $\hat{\beta}_{\mathrm{pw}} = \beta + [\sum_{i\in s} w_i \mathrm{x}_i \mathrm{x}_i']^{-1} \sum_{i\in s} \mathrm{x}_i w_i \varepsilon_i$ where the ε_i 's are the population model residuals. The key result leading to the consistency of $\hat{\beta}_{\mathrm{pw}}$ under the sample distribution is that if $\Pr(I_i = 1 \mid y_i, \mathrm{x}_i, \pi_i) = \pi_i$ then $E_s(w_i \varepsilon_i) = E_s(w_i) E_p(\varepsilon_i) = 0$ (follows from 3.14 below). In fact, by viewing the covariates as random with (y_i, x_i) having some joint distribution,

$$\beta = \arg\min_{\tilde{\beta}} E_p(y_i - \mathrm{x}_i'\tilde{\beta})^2 = \arg\min_{\tilde{\beta}} E_s[w_i(y_i - \mathrm{x}_i'\tilde{\beta})^2],$$

implying that $\hat{\beta}_{\mathrm{pw}}$ is the optimal estimator (in weighted least-squares metric) of β under the sample distribution of (y_i, x_i). See also (3.24) below. Godambe and Thompson (1986, 2009) establish and discuss other optimality properties of estimators solving estimating equations of the form $\sum_{i\in s} w_i u(y_i, \mathrm{x}_i; \theta) = 0$. The following example shows how probability weighting can be used when modelling clustered populations.

Example 7. Consider the population two-level (random intercept) model,

Level 1:

$$u_i \sim N(\mathrm{t}_i'\gamma, \sigma_u^2),\ i = 1,\ldots,M \quad (3.11)$$

Level 2:

$$y_{ij} = \mathrm{x}_{ij}'\beta + u_i + \varepsilon_{ij},\ \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2),\ j = 1 \ldots N_i$$

where ε_{ij} and u_i are independent for all i and j. The unknown parameters are the vectors of coefficients $\vartheta = (\beta', \gamma')'$ and the variances $\tau = (\sigma_\varepsilon^2, \sigma_u^2)'$. Assume full response. Under ignorable sampling of first and second-level units, the *mle* of (ϑ, τ) is computed conveniently by iterating between the estimation of ϑ for 'known' τ and the estimation of τ for 'known' ϑ, with the 'known' values defined by the estimators from the previous iteration. The two sets of estimators on the r^{th} iteration are the solutions of linear equations of the form, $P^{(r)}\vartheta = q^{(r)}$, $R^{(r)}\tau = s^{(r)}$, with appropriate definition of the matrices $(P^{(r)}, R^{(r)})$ and the vectors $(q^{(r)}, s^{(r)})$, $r = 1, 2, \ldots,$ (Goldstein 1986). When applied to all the population values, these equations define the *census* estimating equations.

Suppose, as before, that a sample s_1 of first-level units is sampled with probabilities $\pi_i = \Pr(i \in s_1)$, and that subsamples s_{2i} of size $n_i < N_i$ are sampled from each selected first-level unit i with probabilities $\pi_{j|i} = \Pr(j \in s_{2i} \mid i \in s_1)$. The *pmle* for this model can be obtained by first expressing the elements of the matrices $(P^{(r)}, R^{(r)})$ and the vectors $(q^{(r)}, s^{(r)})$ as sums over first and second-level units, and then estimating each population sum of the form $\sum_{i=1}^{M} d_i$ by the H-T estimator $\sum_{i\in s_1}(d_i/\pi_i)$, and each population sum of the form $\sum_{j=1}^{N_i} d_{ij}$ by the H-T estimator $\sum_{j\in s_{2i}}(d_{ij}/\pi_{j|i})$. See Pfeffermann, Skinner, Holmes, Goldstein and Rasbash (1998b). Pfeffermann and Sverchkov (2009) review other methods of probability weighting in two-level models.

Probability weighting is in broad use both for estimation of finite-population quantities, referred to in the literature as descriptive inference, and for 'analytic inference' on population models. The main attraction of this method is its simplicity. It is generally viewed as being 'model free', except when having to estimate the response probabilities, which is often based on models, and hence more robust than other methods, but when used for analytical inference, this view is questionable.

Probability-weighted estimators are randomization consistent for the corresponding descriptive population quantities (CDPQ), defined as the (hypothetical) solutions of the census estimating equations. However, if the population model is misspecified, the target CDPQ are not (model) p-consistent for the true model parameters and the PW estimators are not $r - p$ consistent either. So, probability weighting provides no protection against model misspecification, although the estimated CDPQ may be useful for various kinds of inference. See Pfeffermann (1993) and Binder and Roberts (2009) for discussion and examples.

Estimating the randomization variance of probability-weighted estimators is generally simple, utilizing available techniques in finite population sampling. Binder (1983) developed a general approach for estimating the randomization variance of estimators obtained as the solution of probability-weighted estimating equations; see also Binder and Roberts (2009) and Godambe and Thompson (2009). Fuller (1975), Binder (1983), Chambless and Boyle (1985) and Francisco and Fuller (1991) developed central limit theorems applicable to probability-weighted estimators.

In spite of these desirable properties of probability-weighting, the method has some severe limitations:

1. It is restricted mostly to point estimation. Probabilistic inference like confidence intervals or

hypothesis testing generally requires large sample normality assumptions. In particular, the randomization distribution does not lend itself to the use of classical inference methods such as likelihood-based or Bayesian inference.

2. The variances of probability-weighted estimators are computed with respect to the randomization distribution and the use of this approach does not permit conditioning on the selected sample, for example, conditioning on the observed covariates or the selected clusters in a multi-level model.
3. As often illustrated in the literature, probability-weighted estimators generally have larger variances than model-based estimators, notably for small samples and large variation of the sampling weights.
4. The use of the randomization distribution does not lend itself to prediction problems such as the prediction of the outcome for non-sampled units with known covariates under a regression model, or the prediction of small area means for areas with no samples in a small-area estimation problem.

3.5 Modifications of the sampling weights

When estimating finite population quantities, the sampling weights are often modified by imposing calibration equations, which match the PW estimators of covariates for which the population totals are known with the actual totals. The use of calibration is particularly useful in the case of nonresponse; see Kott (2009) for recent discussion with references. We later discuss the use of *empirical likelihood* for analytical inference on population models, which also attempts to incorporate calibration equations, although in a different manner. Below, I review two modifications of the sampling weights aimed at reducing the variances of the weighted estimators of model parameters under the *sample distribution* (2.1). A combination of the two modifications is also considered.

Magee (1998) considers a linear regression model but the results can be extended to other population models. The author shows that under certain moment assumptions, any estimator $\hat{\beta}_{\text{mg}}(a) = [\sum_{i\in s} w_i\, a_i(\alpha)\,\mathrm{x}_i\mathrm{x}_i']^{-1}\sum_{i\in s} w_i\, a_i(\alpha)\,\mathrm{x}_i y_i$ with positive weights $a_i(\alpha) = a(\mathrm{x}_i,\, \alpha)$ is consistent for β under the sample distribution. The weights $a(\mathrm{x}_i,\, \alpha)$ belong to a parameterized family of functions with the vector parameter α chosen to minimize a scalar variance criterion such as the determinant or the trace of the asymptotic variance estimator,

$$\begin{aligned} &A\,\hat{\text{var}}[\hat{\beta}_{\text{mg}}(a)] \\ &= \left[\sum\nolimits_{i\in s} w_i\, a_i(\alpha)\mathrm{x}_i\mathrm{x}_i'\right]^{-1} \sum\nolimits_{i\in s} w_i^2 a_i^2(\alpha)\,\hat{\varepsilon}_i^2\mathrm{x}_i\mathrm{x}_i' \\ &\left[\sum\nolimits_{i\in s} w_i\, a_i(\alpha)\mathrm{x}_i\mathrm{x}_i'\right]^{-1}, \end{aligned} \qquad (3.12)$$

where $\hat{\varepsilon}_i = (y_i - \mathrm{x}_i'\hat{\beta}_{\text{pw}})$. The choice of the function $a(\mathrm{x_i},\, \alpha)$ is up to the analyst but the obvious idea is to choose a function that is believed to be approximately inversely proportional to the residual variance under the sample model. The resulting 'Quasi-Aitken' estimator is shown to have asymptotically a lower variance under the sample distribution than the probability-weighted estimator $\hat{\beta}_{\text{pw}}$. Recall from Remark 8 that $\hat{\beta}_{\text{pw}}$ is consistent for β under the sample distribution, justifying comparing the asymptotic variances of the two estimators under this distribution.

Pfeffermann and Sverchkov (1999) propose another modification. Consider the population model,

$$y_j = m(\mathrm{x}_j;\, \theta) + \varepsilon_j,\; E_p(\varepsilon_j \mid \mathrm{x}_j) = 0,\; E_p(\varepsilon_j^2 \mid \mathrm{x}_j) = \sigma^2, \qquad (3.13)$$

where $m(\mathrm{x}_j;\, \theta)$ has a known form. Let $q_i = w_i / E_s(w_i \mid \mathrm{x}_i)$. The authors show that if $\Pr(I_i = 1 \mid \pi_i,\, y_i,\, \mathrm{x}_i) = \pi_i$,

$$E_p(y_i \mid \mathrm{x}_i) = E_s(w_i y_i \mid \mathrm{x}_i) / E_s(w_i \mid \mathrm{x}_i). \qquad (3.14)$$

Thus, for vectors $\tilde{\theta}$ in the plausible parameter space Θ,

$$\begin{aligned} \theta &= \underset{\tilde{\theta}}{\text{argmin}} \frac{1}{n}\sum\nolimits_{i\in s} E_p\{[y_i - m(\mathrm{x}_i;\, \tilde{\theta})]^2 \mid \mathrm{x}_i\} \\ &= \underset{\tilde{\theta}}{\text{argmin}} \frac{1}{n}\sum\nolimits_{i\in s} E_s\{q_i[y_i - m(\mathrm{x}_i;\, \tilde{\theta})]^2 \mid \mathrm{x}_i\}. \end{aligned}$$

The vector θ can be estimated therefore by solving the minimization problem,

$$\begin{aligned} \hat{\theta}_q &= \underset{\tilde{\theta}}{\arg\min} \frac{1}{n}\sum\nolimits_{i=1}^n \hat{q}_i[y_i - m(\mathrm{x}_i;\, \tilde{\theta})]^2; \\ \hat{q}_i &= w_i / \hat{E}_s(w_i \mid \mathrm{x}_i). \end{aligned} \qquad (3.15)$$

The use of this estimator requires estimating $E_s(w_i \mid \mathrm{x}_i)$ but under mild regularity conditions $\hat{\theta}_q$ is consistent for θ even when the expectation $E_s(w_i \mid \mathrm{x}_i)$ is misspecified. See Pfeffermann and Sverchkov (2009) and Section 4.1 of this paper for examples of the specification and estimation of $E_s(w_i \mid \mathrm{x}_i)$.

Example 8. Under the linear regression population model with constant variance,

$$\hat{\beta}_q = \left[\sum_{i\in s} \hat{q}_i \mathrm{x}_i \mathrm{x}_i'\right]^{-1} \sum_{i\in s} \hat{q}_i \mathrm{x}_i y_i. \qquad (3.16)$$

As easily verified, $\hat{\beta}_q$ is randomization consistent for the census regression coefficients $\tilde{B} = [\sum_{j=1}^{N} \mathrm{x}_j \mathrm{x}_j' / E_s(w_j | \mathrm{x}_j)]^{-1} \sum_{j=1}^{N} \mathrm{x}_j \mathrm{y}_j / E_s(w_j | \mathrm{x}_j)$, and hence $p-r$ consistent for β, even when $E_s(w_i | \mathrm{x}_i)$ is misspecified.

The obvious difference between the PW estimator $\hat{\theta}_{\mathrm{pw}}$ and the estimator $\hat{\theta}_q$ is that the latter estimator uses the adjusted weights $q_i = w_i / \hat{E}_s(w_i | \mathrm{x}_i)$. When the sample selection depends only on the covariates, the sampling process is ignorable. Hence, to protect against informative sampling, it is only necessary to account for the net sampling effects on the target conditional *pdf* of $y_i | \mathrm{x}_i$. This is achieved by using the weights q_i. In contrast, the sampling weights w_i account for the sampling effects on the joint distribution of (y_i, x_i). As a result, they tend to be more variable and the estimator $\hat{\theta}_{\mathrm{pw}}$ has a larger variance.

A combination of the last two modifications is also possible and examined in Section 4. The simple idea proposed by Dr. Moshe Feder (private communication) is to apply the modification of Magee (1998) to the estimator $\hat{\beta}_q$ instead of the estimator $\hat{\beta}_{\mathrm{pw}}$, that is, use the estimator,

$$\hat{\beta}_{\mathrm{mg}-q}(a) =$$

$$\left[\sum\nolimits_{i\in s} \hat{q}_i\, a_{i,q}(\alpha) \mathrm{x}_i \mathrm{x}_i'\right]^{-1} \sum\nolimits_{i\in s} \hat{q}_i\, a_{i,q}(\alpha) \mathrm{x}_i y_i, \qquad (3.17)$$

where the vector parameter α is now chosen to minimize a scalar variance criterion of the asymptotic variance estimator, $A\hat{\mathrm{var}}[\hat{\beta}_{\mathrm{mg}-q}(a)]$, computed similarly to (3.12).

3.6 Likelihood based methods

3.6.1 Use of the sample model for maximum likelihood estimation

A natural way of estimating the population model parameters is by maximization of the sample likelihood. Assume first full response and that the sample observations are independent under the sample distribution. The likelihood has then the form,

$$L_s(\theta, \gamma; y_s, \mathrm{x}_s)$$

$$= \prod\nolimits_{i\in s} \frac{\Pr(I_i = 1 \mid \mathrm{x}_i, y_i; \gamma)\, f_p(y_i \mid \mathrm{x}_i; \theta)}{\Pr(I_i = 1 \mid \mathrm{x}_i; \gamma, \theta)}. \qquad (3.18)$$

As before, we assume $\Pr(I_i = 1 \mid \pi_i, y_i, \mathrm{x}_i) = \pi_i$, implying $\Pr(I_i = 1 \mid \mathrm{x}_i, y_i) = E_p(\pi_i \mid \mathrm{x}_i, y_i)$. By (3.14), The sample likelihood can be written therefore as,

$$L_s(\theta, \gamma; y_s, \mathrm{x}_s) = \prod\nolimits_{i\in s} \frac{E_s(w_i | \mathrm{x}_i; \theta, \gamma) f_p(y_i | \mathrm{x}_i; \theta)}{E_s(w_i \mid y_i, \mathrm{x}_i; \gamma)}. \quad (3.19)$$

The expectations on the right hand side of (3.19) are with respect to the sample *pdf* of the sampling weights. Thus, when the weights are known for the sampled units as is usually the case under full response, the expectations can be modelled and estimated by regressing w_i against (y_i, x_i), using classical model fitting procedures. Suppose first that the weights are continuous such as in probability proportional to size (PPS) sampling with a continuous size variable. For a given form of the population model, the expectations $E_s(w_i \mid y_i, \mathrm{x}_i; \gamma)$ and $E_s(w_i \mid \mathrm{x}_i; \gamma, \theta)$ can be obtained then in two steps:

1. Identify and estimate $\hat{E}_s(w_i | y_i, \mathrm{x}_i; \gamma) = E_s(w_i | y_i, \mathrm{x}_i; \hat{\gamma})$, using the sample data.
2. Integrate $\int [1/E_s(w_i | y, \mathrm{x}_i; \hat{\gamma})] f_p(y \mid \mathrm{x}_i; \theta) dy$ to obtain $E_p(\pi_i | \mathrm{x}_i; \theta; \hat{\gamma})$. Compute, $\hat{E}_s(w_i | \mathrm{x}_i; \theta, \hat{\gamma}) = 1/E_p(\pi_i | \mathrm{x}_i; \theta, \hat{\gamma})$ (follows from 3.14).

Estimating the vector parameter γ outside the likelihood and then substituting the estimate in (3.19) and maximizing the likelihood as a function of the vector parameter θ only, usually yields more stable results than maximizing the likelihood over (θ, γ) simultaneously.

Estimation of the expectations $E_s(w_i \mid y_i, \mathrm{x}_i; \gamma)$ and $E_s(w_i \mid \mathrm{x}_i; \theta, \gamma)$ in the case of discrete inclusion probabilities is similar.

Example 9. Consider the case of multinomial-logistic regression with a discrete covariate x and M possible values of the outcome y. Assuming that $E_s(w_i \mid y_i = m, x_i = k)$ is not a function of the model parameters, it can be estimated by $\bar{w}_{mk}$, the mean of the weights in cell (m, k), and thence $\hat{\pi}_{mk} = \hat{\Pr}_p(i \in s \mid y_i = m, x_i = k) = (1 / \bar{w}_{mk})$. We obtain:

$$\Pr_s(y_i = m \mid x_i = k; \theta)$$

$$\cong \frac{[\Pr_p(y_i = m \mid x_i = k; \theta) / \bar{w}_{mk}]}{\sum_{m^*=1}^{M} [\Pr_p(y_i = m^* \mid x_i = k; \theta) / \bar{w}_{m^*k}]}. \qquad (3.20)$$

The sampling weights feature in the sample model, but this is not an application of classical probability weighting. Notice that with this approximation the parameters in the population and the sample model are the same. In our empirical study we use a similar approximation for the sample distribution by categorizing the values of a continuous outcome. See Pfeffermann and Sverchkov (1999) for other examples.

Next consider the estimation of the vector parameter θ governing the population model. Under mild conditions, θ is the unique solution of the equations,

$$W_U(\theta) = \sum\nolimits_{j\in U} E_p(\delta_j \mid \mathrm{x}_j) = 0;$$

$$\delta_j = (\delta_{j,0}, \delta_{j,1}, \ldots, \delta_{j,k})' = \partial \log f_p(y_j \mid \mathrm{x}_j; \theta) / \partial\theta. \quad (3.21)$$

Pfeffermann and Sverchkov (2003) consider three different approaches for estimating θ. The common feature of these approaches is that the only data used for estimation are the observations $\{(y_i, \mathrm{x}_i, w_i), i \in s\}$, similarly to the PW estimators and their modifications considered in Section 3.5. In Section 3.6.2 we consider the use of the 'full likelihood', which assumes knowledge of the covariates $\{\mathrm{x}_j, j \in U\}$, and possibly also additional design information.

The first approach redefines the parameter equations with respect to the sample model. Assuming that $E_s(w_i \mid \mathrm{x}_i; \theta, \gamma)$ in (3.19) is differentiable with respect to θ, the sample model parameter equations are $W_{1s}(\theta) = \sum_{i \in s} E_s \{[\partial \log f_s(y_i \mid \mathrm{x}_i; \theta, \gamma) / \partial\theta] \mid \mathrm{x}_i\} = \sum_{i \in s} E_s \{[\delta_i + \partial \log E_s(w_i \mid \mathrm{x}_i; \theta, \gamma) / \partial\theta] \mid \mathrm{x}_i\} = 0$. The vector θ is estimated under this approach by solving the equations,

$$W_{1s,e}(\theta) = \sum_{i \in s} [\delta_i + \partial \log E_s(w_i \mid \mathrm{x}_i; \theta, \gamma) / \partial\theta] = 0. \quad (3.22)$$

The second approach applies the relationship (3.14) to the parameter equations (3.21). For a random sample from the sample model, the equations are now $W_{2s}(\theta) = \sum_{i \in s} E_s(q_i \delta_i \mid \mathrm{x}_i) = 0$, where $q_i = w_i / E_s(w_i \mid \mathrm{x}_i)$. The vector θ is estimated under this approach by solving the equations,

$$W_{2s,e}(\theta) = \sum_{i \in s} q_i \delta_i = 0. \quad (3.23)$$

The third approach uses the property that if θ solves (3.21), then it solves also the equations, $\tilde{W}_U(\theta) = \sum_{j \in U} E_p(\delta_j) = E_{\mathrm{x}}[\sum_{j \in U} E_p(\delta_j \mid \mathrm{x}_j)] = 0$, where $E_{\mathrm{x}}(\cdot)$ is the expectation of x (which is viewed as random) with respect to the population distribution. Hence, by (3.14), for a random sample from the sample model, the parameter equations are $W_{3s}(\theta) = \sum_{i \in s} E_s(w_i \delta_i) = 0$, with estimating equations,

$$W_{3s,e}(\theta) = \sum_{i \in s} w_i \delta_i = 0. \quad (3.24)$$

Note that the equations (3.24) are the *pseudo-likelihood* equations (Remark 7).

Remark 9. The use of the weights $q_i = w_i / E_s(w_i \mid \mathrm{x}_i)$ for population model parameter estimation has been justified already in Section 3.5 by reference to least-squares estimation. See the discussion in that section regarding the difference between the use of the weights q_i and the weights w_i. Pfeffermann and Sverchkov (1999, 2003) illustrate that estimating θ by solving the equations (3.23) yields estimators with lower randomization variance than estimating θ by solving the equations (3.24). Notice that under the assumption of a linear regression model operating in the population, the solution of (3.24) yields the PW estimator (3.8), and the solution of (3.23) yields the q-weighted estimator (3.16).

Remark 10. The use of the sample model for estimation of multi-level population models is considered in Pfeffermann, Moura and Nascimento-Silva (2006), using the Bayesian approach. Pfeffermann and Sverchkov (2007) fit multi-level models for small area estimation under informative sampling of areas and within the areas, following the frequentist approach.

So far we assumed full response. Next consider the case of NMAR nonresponse. In this case the response process needs to be modelled as well. By (2.2) and with added parameter notation the 'respondents' likelihood takes the form,

$$L_o = \prod_{i=1}^{r} f(y_i \mid \mathrm{x}_i, I_i = 1, R_i = 1; \theta^*, \gamma^*)$$

$$= \prod_{i=1}^{r} \frac{\Pr(R_i = 1 \mid y_i, \mathrm{x}_i, I_i = 1; \gamma^*) f_s(y_i \mid x_i; \theta^*)}{\Pr(R_i = 1 \mid \mathrm{x}_i, I_i = 1; \gamma^*, \theta^*)}, \quad (3.25)$$

where $\theta^* = (\theta, \gamma)$ represents the parameters of the sample distribution under full response (Equation 3.19), and γ^* represents the parameters of the response process. Notice that unlike the sampling probabilities $\pi_i = \Pr(i \in s)$, which are generally known and can be used for estimating the probabilities $\Pr(I_i = 1 \mid y_i, \mathrm{x}_i; \gamma)$ as explained before, the response probabilities are generally unknown.

Chang and Kott (2008) propose a method of estimating the response probabilities, which uses known totals of calibration variables. The authors assume a parametric model for the response probabilities that may depend on the outcome value, and estimate the unknown parameters of this model by regressing the totals of the calibration variables against their H-T estimators. The weights used for the H-T estimators are the product of the sampling weights and the inverse of the response probabilities under the model. Let c_i define the values of the calibration variables for unit i and denote $p(y_i, \mathrm{x}_i; \gamma^*) = \Pr(R_i = 1 \mid y_i, \mathrm{x}_i, I_i = 1; \gamma^*)$. Chang and Kott (2008) estimate the unknown parameters by setting the nonlinear regression equations,

$$C^U = \sum_{i=1}^{r} w_i \frac{c_i}{p(y_i, \mathrm{x}_i; \gamma^*)} + \varepsilon^*,$$

where $C^U = \sum_{j=1}^{N} c_j$ and ε^* is a vector of errors. The parameters γ^* are estimated by the iterative algorithm

$$\hat{\gamma}^{(j+1)} = \hat{\gamma}^{(j)} + \left\{\hat{H}(\hat{\gamma}^{(j)})^T V^{-1}(\hat{\gamma}^{(j)}) \hat{H}(\hat{\gamma}^{(j)})\right\}^{-1}$$

$$\hat{H}(\hat{\gamma}^{(j)})^T V^{-1}(\hat{\gamma}^{(j)}) \left(C^U - \sum_{i=1}^{r} w_i \frac{c_i}{\pi(y_i, v_i; \hat{\gamma}^{(j)})} \right), \quad (3.26)$$

where

$$\hat{H}(\hat{\gamma}^{(j)}) = \frac{\partial \left[\sum_{i=1}^{r} w_i \frac{c_i}{\pi(y_i, v_i; \gamma)} \right]}{\partial\gamma} \Bigg|\; \gamma = \hat{\gamma}^{(j)} \text{ and } V^{-1}(\hat{\gamma}^{(j)})$$

is the inverse of the estimated quasi-randomization variance of

$$\sum_{i=1}^{r} w_i \frac{c_i}{\pi(y_i, v_i; \gamma)},$$

computed at $\gamma = \hat{\gamma}^{(j)}$.

Chang and Kott (2008) do not assume a model for the outcome and their approach is therefore restricted to estimation of the model for the response probabilities. Pfeffermann and Sikov (2011) use the likelihood (3.25) for estimating population models assuming noninformative sampling. Maximization of the likelihood is carried out by iterating between maximization of the likelihood with respect θ^* for given γ^*, and the solution of calibration equations with respect to γ^* for given θ^*, using known totals of calibration variables, similarly to Chang and Kott (2008). The 'given' parameters are the estimates from the previous iteration. The authors show how to estimate the distribution of the missing covariates and outcome for a nonresponding unit and use this distribution for imputing the missing outcomes and hence predicting the finite population total of the outcome variable.

Estimation of the population model by fitting the sample model has some important advantages not shared by the other approaches considered in this article.

1. Once the sample model is specified, it lends itself to standard model based inference such as likelihood based methods, Bayesian inference or semi-parametric modelling. It is important to emphasize in this regard that the goodness of fit of the postulated population model can be evaluated by testing the goodness of fit of the sample model fitted to the observed outcomes, using classical model diagnostic techniques. See Krieger and Pfeffermann (1997) and Pfeffermann and Sikov (2011) for appropriate test statistics with illustrations.
2. The sample likelihood provides a coherent way of handling NMAR nonresponse when estimating population models. Methods based on probability weighting require knowledge or good estimators of the response probabilities. The use of the full likelihood (see below) requires knowledge of the covariates of nonsampled units.
3. Application of this approach permits the use of conditional inference, given the sample of responding units, for example, conditioning on the observed covariates.
4. The models holding for the observed outcomes and the response probabilities define the model holding for the missing outcomes of the non-sampled units or the nonrespondents, which can be used for imputation of these outcomes. Methods based on probability weighting and variants thereof allow estimating the population model but under informative sampling and NMAR nonresponse, the population model cannot be used for prediction or imputation of the missing outcomes. See Sverchkov and Pfeffermann (2004) and Pfeffermann and Sikov (2011) for illustrations.
5. The use of the sample model enables testing whether the sampling process can be ignored. Pfeffermann and Sverchkov (2009) review several test statistics proposed in the literature for testing the ignorability of the sample selection.

3.6.2 The full likelihood

Theoretically, a more efficient way of estimating the unknown population model parameters is to base the likelihood on the joint distribution of the sample data and the sample membership indicators. Under full response, the *full likelihood* is then,

$$L_f(\theta, \gamma; \mathrm{I}_U, \mathrm{y}_s, \mathrm{x}_s, \mathrm{x}_{\tilde{s}}) =$$

$$\prod_{i \in s} \Pr(I_i = 1 \mid y_i, \mathrm{x}_i; \gamma) f_p(y_i \mid \mathrm{x}_i; \theta)$$

$$\prod_{j \notin s} [1 - \Pr(I_j = 1 \mid \mathrm{x}_j; \theta, \gamma)], \qquad (3.27)$$

where $\mathrm{I}_U = \{I_1, \ldots, I_N\}$ is the vector of sample inclusion indicators and $\Pr(I_j = 1 \mid \mathrm{x}_j;\, \theta, \gamma) = \int \Pr(I_j = 1 \mid y_j, \mathrm{x}_j, \gamma) f_p(y_j \mid \mathrm{x}_j, \theta)\, dy_j$ is the *propensity score* of unit *j*. The likelihood (3.27) assumes $\Pr(\mathrm{I}_U \mid \mathrm{y}_U, \mathrm{x}_U) = \prod_{k \in U} \Pr(I_k \mid y_k, \mathrm{x}_\mathrm{k})$ (Poisson sampling), but it can be generalized to other sampling designs. The full likelihood has the advantage of accounting for the sampling probabilities of units outside the sample, thus utilizing more information, but it requires knowledge of the covariates of all the population units. See, for example, Gelman, Carlin, Stern and Rubin (2003) and Little (2004). Modelling the joint distribution of the covariates for units outside the sample and integrating them out of the likelihood can be very complicated in practice and is formidable when there are many of them. Pfeffermann *et al.* (2006) compare empirically the use of the sample likelihood with the use of the full likelihood for multi-level models in a Bayesian context. The two approaches yield similar results, but this of course may not be the case in other applications.

Another way of defining the full likelihood is by application of the *Missing Information* Principle (MIP, Orchard and Woodbury 1972). The basic idea is to express the sample score function as the conditional expectation of the population score function, given the sample data. Following Chambers and Skinner (2003, Chapter 2), define the *full-sample likelihood* as $L_{fs}(\lambda) = f(\lambda; y_s, \mathrm{x}_s, I_U, z_U)$

where, as before, z_U is a known matrix of population values underlying the sample selection and λ defines the unknown model parameters. The corresponding *full-population* likelihood is $L_{fU}(\lambda) = f(\lambda; \mathrm{y}_U, \mathrm{x}_U, I_U, \mathrm{z}_U)$ where $\mathrm{y}_U = (\mathrm{y}_s, \mathrm{y}_{\bar{s}})$ and $\mathrm{x}_U = (\mathrm{x}_s, \mathrm{x}_{\bar{s}})$. The MIP principle states that,

$$\begin{aligned} sc_s(\lambda) &= (\partial / \partial\lambda)\log[L_{fs}(\lambda)] \\ &= E_p[(\partial / \partial\lambda)\log L_{fU}(\lambda) \mid y_s, \mathrm{x}_s, I_U, \mathrm{z}_U]. \quad (3.28)\end{aligned}$$

Another identity defines the relationship between the population likelihood information matrix and the sample likelihood information matrix.

Breckling, Chambers, Dorfman, Tam and Welsh (1994) and Chambers *et al.* (1998) consider applications of the MIP to complex survey data. In particular, Chambers *et al.* (1998) study the use of the MIP when only limited design information is available and not the full information entailed in z_U. The authors show examples where the use of the MIP is more efficient than the use of the sample likelihood $L_s(\theta, \gamma; y_s, \mathrm{x}_s)$ defined by (3.19), which only uses the weights $\{w_i,\ i \in s\}$. The likelihood (3.28) can be extended to account for NMAR nonresponse but the application of this approach requires then knowledge of the population values of the variables explaining the response. The computation of the expectation in the right hand side of (3.29) may not be simple either, depending on the population model.

Remark 11. The use of the MIP method in the simulation set up of Section (3.1) requires knowledge of the covariates and stratification membership for units outside the sample. We didn't find a way of applying the method in this case without further assumptions on the joint distribution of the covariates and the design variables.

3.6.3 Empirical likelihood

In recent years there is a growing interest in the use of empirical likelihood (EL) methods for analyzing complex survey data. The EL method as originally proposed by Hartley and Rao (1968) in the survey sample context and by Owen (1988, 2001) combines the robustness of nonparametric methods with the effectiveness of the likelihood approach. Two other important advantages of this method are that it lends itself very naturally to the use of calibration equations and that it enables the construction of confidence intervals without the need for variance estimation.

Consider the model defined by (3.13) where for now we view the covariates as random, and denote $\mathrm{g}_i = (y_i, \mathrm{x}_i')'$. Under some regularity conditions, the vector parameter θ is the unique solution of the equation

$$E_p\left\{\frac{\partial m(\mathrm{x};\, \theta)}{\partial\theta}[y - m(\mathrm{x};\, \theta)]\right\} = 0.$$

Let $p_1, \ldots, p_n$ be a set of probabilities corresponding to the observations $(\mathrm{g}_1, \ldots, \mathrm{g}_n)$ such that p_i is the 'jump' (probability mass) of the population cumulative distribution $F_p(\mathrm{g}_i)$ at g_i. It is assumed that F_p has its support on the observed values such that

$$\textstyle\sum_{i=1}^n p_i \dfrac{\partial m(\mathrm{x}_i;\, \theta)}{\partial\theta}[y_i - m(\mathrm{x}_i;\, \theta)] = 0. \quad (3.29)$$

Assuming independent observations, the EL of F_p is $L(F_p) = \prod_{i=1}^n p_i$. Notice that if p_i is a known function of some unknown parameters, $L(F_p)$ coincides with the standard parametric likelihood. The (nonparametric) EL estimators of the probabilities p_i are the solution $p_i^{(p)}$ of the maximization problem,

$$\max_{p_1,\ldots,p_n} \prod_{i=1}^n p_i \ \text{ s.t. } p_i \geq 0, \textstyle\sum_{i=1}^n p_i = 1, \quad (3.30)$$

yielding $p_i^{(p)} = 1/n, i = 1, \ldots, n$. For the linear regression case, $m(\mathrm{x}_i;\, \theta) = \mathrm{x}_i'\beta$ and by substituting $p_i^{(p)}$ for p_i in (3.29) and solving the equations we obtain the EL estimator of β as $\hat{\beta}_{\mathrm{el}} = \hat{\beta}_{\mathrm{OLS}}$. When finite population means $\bar{\mathrm{C}}^U$ of variables C measured in the sample are known, they can be added to the maximization problem (3.30) by adding the calibration constraints $\sum_{i=1}^n p_i \mathrm{c}_i = \bar{\mathrm{C}}^U$. This additional information is expected to enhance the estimation of the p_i's and hence the estimation of the unknown model parameters. See also Remark 12 below.

Suppose now that units are drawn to the sample (or respond) with unequal selection probabilities π_i. In this case it is common to replace the objective empirical likelihood $L(F_p) = \prod_{i=1}^n p_i$ by the pseudo empirical likelihood $L_{\mathrm{pl}}(F_p) = \prod_{i=1}^n p_i^{w_i}$, where, as before, $w_i = 1/\pi_i$. Notice that $\log L_{\mathrm{pl}}(F_p) = \sum_{i=1}^n w_i \log(p_i)$ is the H-T estimator of $\log L_{\mathrm{pop}}(F_p) = \sum_{i=1}^N \log p_i$. The pseudo EL estimators of the p_i's solve the maximization problem,

$$\max_{p_1,\ldots,p_n} = \prod_{i=1}^n p_i^{w_i} \ \text{ s.t. } p_i \geq 0, \textstyle\sum_{i=1}^n p_i = 1. \quad (3.31)$$

See, *e.g.*, Chen and Sitter (1999). It is easy to verify that in the absence of benchmark constraints, the solution of (3.31) is $p_i^{(\mathrm{pel})} = w_i / \sum_{i=1}^n w_i$ and by substituting $p_i^{(\mathrm{pel})}$ for p_i in (3.29), $\hat{\beta}_{\mathrm{pel}} = \hat{\beta}_{\mathrm{pw}}$, the PW estimator (3.8).

The empirical likelihoods in (3.30) and (3.31) are with respect to the population distribution. Alternatively, one can obtain the EL estimator by defining the likelihood with respect to the sample distribution $f_s(\mathrm{g}_i) = \Pr(I_i = 1 \mid \mathrm{g}_i)$ $f_p(\mathrm{g}_i) / \Pr(I_i = 1)$, where by denoting $\tau_i = \Pr(I_i = 1 \mid \mathrm{g}_i)$, $\Pr(I_i = 1) = \sum_{i=1}^n p_i \tau_i$. Following Kim (2009) and Chaudhuri, Handcock and Rendall (2010), the EL estimators of the probabilities p_i are obtained now as the solution of the maximization problem

$$\max_{p_1,\ldots,p_n}\left[\sum\nolimits_{i=1}^{n}\log(p_i\tau_i) - n\log\sum\nolimits_{i=1}^{n}p_i\tau_i\right]$$

$$\text{s.t. } p_i \geq 0, \sum\nolimits_{i=1}^{n}p_i = 1. \qquad (3.32)$$

The solution of (3.32) is $p_i^{\text{sel}} = \tau_i^{-1} / \sum_{j=1}^{n}\tau_j^{-1}$ and by substituting in (3.29),

$$\hat{\beta}_{\text{sel}} = \left[\sum\nolimits_{i=1}^{n}\tau_i^{-1}\mathrm{x}_i\mathrm{x}_i'\right]^{-1}\sum\nolimits_{i=1}^{n}\tau_i^{-1}\mathrm{x}_i\mathrm{y}_i. \qquad (3.33)$$

The estimator $\hat{\beta}_{\text{sel}}$ has the same form as the PW estimator $\hat{\beta}_{\text{pw}}$ in (3.8), but with the weights $\tau_i^{-1} = 1/\Pr(i \in s \mid y_i, \mathrm{x}_i)$ instead of the sampling weights w_i. In practice, one has to replace the probabilities τ_i by sample estimates $\hat{\tau}_i$. See Section 4.

Remark 12. The following possible enhancement to the estimation of the probabilities p_i was proposed to me by Dr. Jae Kim in a private communication. Assuming as before that $\Pr(i \in s \mid \pi_i, y_i, \mathrm{x}_i) = \pi_i$, it follows that $\tau_i = \Pr(I_i = 1 \mid y_i, \mathrm{x}_i) = E_p(\pi_i \mid y_i, \mathrm{x}_i)$ and hence that $E_p[(\pi_i - \tau_i) \mid y_i, \mathrm{x}_i] = 0$. This suggests adding calibration constraints of the form

$$\sum\nolimits_{j=1}^{n}p_j(\pi_j - \hat{\tau}_j)\,k(y_j, \mathrm{x}_j) = 0 \qquad (3.34)$$

to enhance the estimation of the probabilities $\{p_i\}$ in (3.31), where $k(y_j, \mathrm{x}_j) = k(g_j)$ is some function of the observed outcome and covariates. Examples for plausible functions for the case of a single covariate x are, $k(g_j) = y_j x_j$, $k(g_j) = y_j / x_j$ *etc*. The notable feature of the constraints (3.34) is that they do not require knowledge of population quantities like means of calibration variables, as is often assumed when advocating the EL approach for sample survey estimation. Clearly, when means $\bar{C}^U$ of calibration variables are known, constraints of the form $\sum_{i=1}^{n}p_i\mathrm{c}_i = \bar{\mathrm{C}}^U$ may be added as well. See also Remark 14.

4. Empirical study

In this section I report the results of a simulation study aimed at assessing and comparing the performance of the methods discussed in Section 3. The simulation set up is described in Section 3.1 and we use $H = 5$ strata. The target parameters are the regression coefficients $\beta' = (\beta_0, \beta_1) = (2,1)$ of the population expectation (3.1). The simulation study consists of generating 2,000 populations and samples (one sample from each population) and computing the estimators, variance estimators and confidence intervals listed below for each sample. The population size is 5,000 with approximate strata sizes $N_h = 363,\ 554,\ 842,\ 1{,}278,\ 1{,}963$. (The strata sizes are random). The sample size is $n = 300$ with $n_h = 60$ sampled units in each stratum. The sampling fractions are therefore highly variable across the strata.

We generated population values of a single discrete covariate x by first generating observations $\tilde{x}_j$ from a *Gamma* distribution with mean 2 and variance 4, and then defining x_j to be the nearest integer to $\tilde{x}_j$ if $\tilde{x}_j < 5$ and $x_j = 5$ otherwise. The covariates are therefore $\mathrm{x}_j = (1, x_j)'$, with $x_j = 0, 1, \ldots, 5$. The population covariates were generated once and held fixed for all the populations.

Figure 1 shows the population and sample *pdf*s of the outcome y for x = 2, 3, 4, 5.

As can be seen, the population and sample *pdf*s differ, indicating the informativeness of the sampling process. Notice also that the population *pdf* is not normal because the random coefficients ζ_j are not normal.

We study the performance of the various methods in terms of *bias*, *variance*, *variance estimation*, and *confidence interval* coverage. We assume for all the methods that the only available information are the observed outcomes and covariates $(\mathrm{y}_{hs}, \mathrm{x}_{hs})$ for every stratum h, the sample selection probabilities and the true strata sizes $\{N_h\}$. I believe that this is the practice in most real life applications.

4.1 Estimators considered

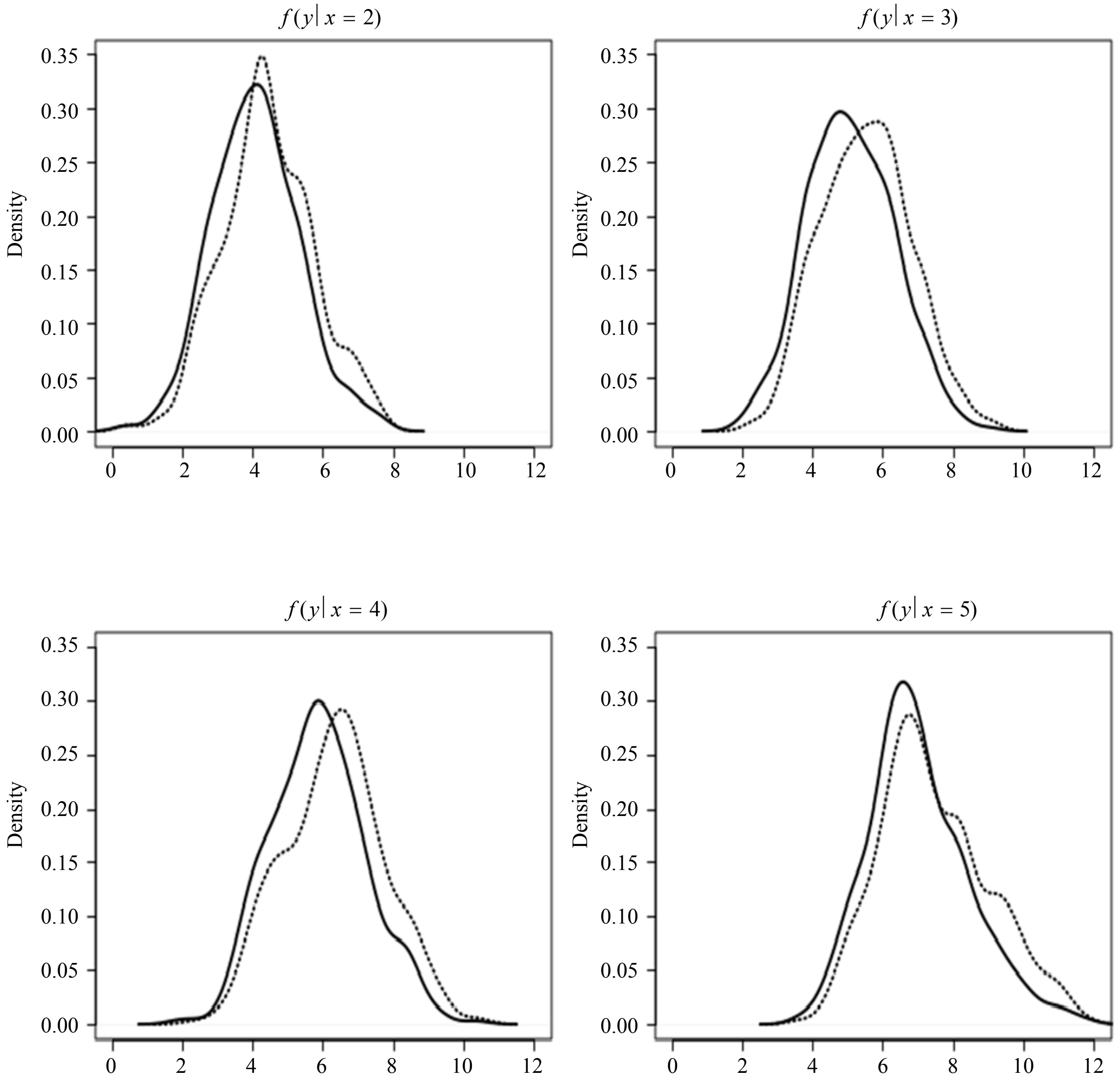
4.1.1 The OLS estimator $\hat{\beta}_{\text{ols}}$. The use of this estimator ignores the sampling process.

4.1.2 The estimator proposed by Feder (2011, see Section 3.2). Application of this approach is in four steps. *i*) fit a linear model with constant residual variance in each stratum, *ii*) impute the missing covariate values for the non-sampled units by sampling with replacement $(N_h - n_h)$ values from the n_h observed values in stratum h with probabilities $\vec{p}_{hi} = (w_{hi} - 1)/\sum_{k=1}^{n_h}(w_{hk} - 1)$ on each draw, where the w_{hi}'s are the sampling weights when sampling from stratum h. *iii*) impute the missing y-values in each stratum by generating observations at random from the model fitted in Step *i*). *iv*) fit the linear regression model of y on x by using all the population data, with the missing values for the non-sampled units replaced by the imputed values. We denote the resulting estimator by $\hat{\beta}_f$.

4.1.3 The PW estimator $\hat{\beta}_{\text{pw}}$ (Equation 3.8).

4.1.4 The estimator $\hat{\beta}_{\text{mg}}$ proposed by Magee (1998, see Section 3.5). In our application we define $a_i(\alpha) = (x_i + 0.1)^{\alpha}$ and search for the optimal power α in the range $[-2, 2]$ minimizing the determinant of the asymptotic variance estimator (3.12).

4.1.5 The estimator $\hat{\beta}_q$ defined by (3.16). For the present study we do not assume any parametric model for the expectation $E_s(w_i \mid \mathrm{x}_i)$ in the denominator of q_i and estimate $\hat{E}_s(w_i \mid \mathrm{x}_i) = \bar{w}_s(\mathrm{x}_i)$, the mean of the observed sampling weights for units with $\mathrm{x} = \mathrm{x}_i$.



Figue 1 Population *pdf* (solid line) and sample *pdf* (dashed line) of *y*|x

4.1.6 The modified q-weighted estimator $\hat{\beta}_{\text{mg}-q}$ defined by (3.17). The weights $\hat{q}_i$ are obtained as in 4.1.5 and the functions $a_{i,q}(\alpha)$ as in 4.1.4.

4.1.7 Estimators derived by maximization of the sample likelihood (3.19). The use of this approach requires specifying the population *pdf* and the expectation $E_s(w_i | y_i, \mathrm{x}_i)$. The unknown population model parameters are $\theta' = (\beta', \sigma^2)$ and we assume $f_p(y_i | \mathrm{x}_i; \theta) = N(\mathrm{x}_i'\beta, \sigma^2)$, which as noted before and illustrated in Figure 1 is not the correct *pdf* since the random coefficients ζ_j are not normal (see Section 3.1). We estimated $E_s(w_i | y_i, \mathrm{x}_i; \gamma)$ nonparametrically and set up the likelihood as follows:

Let s_{x_i} define the sample of units with $\mathrm{x} = \mathrm{x}_i$ of size m_{x_i}. We first divided the sample into $c(\mathrm{x}_i)$ homogeneous clusters based on the ascending values of the outcome y using the R function "hclust". The $c(\mathrm{x}_i)$'s are between 1 and 7, depending on the sample size m_{x_i} (one cluster if $m_{\mathrm{x}_i} \leq 10$, 2 clusters if $m_{\mathrm{x}_i} \leq 20$, ..., 7 clusters if $m_{\mathrm{x}_i} \geq 70$). Denote by $b_{\mathrm{x}_i, k}$ the midpoint between the highest y-value in cluster k and the lowest y-value in cluster $(k+1)$, $k = 1, ..., c(\mathrm{x}_i) - 1$, and define $b_{\mathrm{x}_i,0} = -\infty$, $b_{\mathrm{x}_i,c(\mathrm{x}_i)} = +\infty$. For $b_{\mathrm{x}_i,k-1} \leq y \leq b_{\mathrm{x}_i,k}$ we estimated $E_s(w_i | y_i, \mathrm{x}_i)$ by the mean $\bar{w}_s(y, \mathrm{x}_i) = \bar{w}_k(\mathrm{x}_i)$ of the sampling weights of units with y-values in the same interval. Substituting $E_s(w_i | y_i, \mathrm{x}_i) = \bar{w}_s(y_i, \mathrm{x}_i)$ in (3.19) defines the sample likelihood used for the present simulation study as,

$$L_s(\theta; y_s, \mathrm{x}_s) = \prod_{i \in s} \frac{f_p(y_i \mid \mathrm{x}_i; \theta) \,/\, \overline{w}_s(y_i, \mathrm{x}_i)}{\sum_{k=1}^{c(\mathrm{x}_i)} [F_p(b_{k,\mathrm{x}_i}) - F_p(b_{k-1,\mathrm{x}_i})] \,/\, \overline{w}_k(\mathrm{x}_i)}, \quad (4.1)$$

where $F_p(b_{k,\mathrm{x}_i}) = \int_{-\infty}^{b_k} f_p(y \mid \mathrm{x}_i; \theta)\, dy$ (the CDF of the assumed normal *pdf*).

The approximation (4.1) is similar to the approximation (3.20) proposed for the case where both x and y are discrete.

Remark 13. In order to facilitate the numerical optimizations used for the computation of the estimators $\hat{\beta}_{\mathrm{mg}}, \hat{\beta}_{\mathrm{mg}-q}$ and the maximum likelihood estimators in (4.1), we transformed the minimization problem $\min\{f(\theta): \theta \in (a, b)\}$ to $\min\{f[g(\eta)]: \eta \in (-\infty, \infty)\}$ with the function $g(\eta)$ defined as $g(\eta) = [(b-a)\tan^{-1}(\eta)] / \pi + 0.5(a+b)$. Notice that every $\theta \in (a, b)$ has an image $\eta \in R;\ g(\eta) = \theta$, and $\arg\min\{f(\theta): \theta \in (a, b)\} = g(\eta_0)$ where $\eta_0 = \arg\min f[g(\eta)]$.

We used the R function *nlm* for the numerical optimization, with the PW estimates as starting values. To prevent numerical overflows of the optimized function by evaluation of exponentials of large numbers, the maximization was limited to the intervals $\{\min[0.5\hat{\beta}_{\mathrm{pw}}, \hat{\beta}_{\mathrm{pw}} - 3\hat{\mathrm{se}}(\hat{\beta}_{\mathrm{pw}})], \max[1.5\hat{\beta}_{\mathrm{pw}}, \hat{\beta}_{\mathrm{pw}} + 3\hat{\mathrm{se}}(\hat{\beta}_{\mathrm{pw}})]\}$ for β, and $[0.5\hat{\sigma}_{\mathrm{pw}}, 1.5\hat{\sigma}_{\mathrm{pw}}]$ for σ.

4.1.8 The empirical likelihood estimator $\hat{\beta}_{\mathrm{sel}}$ defined by (3.33). The computation of this estimator requires estimating the probabilities $\tau_i = \Pr(I_i = 1 \mid \mathrm{y}_i, \mathrm{x}_i) = 1/E_s(w_i \mid \mathrm{y}_i, \mathrm{x}_i)$, and we use the estimator $\hat{E}_s(w_i \mid y_i, \mathrm{x}_i) = \overline{w}_{s,k}(y, \mathrm{x}_i)$ used for defining the likelihood (4.1), such that $\hat{\tau}_i = 1/\overline{w}_k(y, \mathrm{x}_i)$.

4.2 Variance estimation

We applied three approaches for variance estimation. The first approach estimates the randomization variance, the second approach estimates the variance under the sample model, while the third approach uses the nonparametric bootstrap method, which likewise estimates the variance under the sample model.

Consider first the estimators defined by 4.1.1, 4.1.3 – 4.1.6 and 4.1.8 in Section 4.1. All these estimators can be written in the generic form,

$$\hat{\beta}_t = \left[\sum_{i=1}^{n} w_i t_i \mathrm{x}_i \mathrm{x}_i'\right]^{-1} \sum_{i=1}^{n} w_i t_i \mathrm{x}_i \mathrm{y}_i$$
$$= [X_s' W_s T_s X_s]^{-1} \sum_{i=1}^{n} w_i t_i \mathrm{x}_i \mathrm{y}_i, \quad (4.2)$$

where $X_s' = [\mathrm{x}_1, \ldots, \mathrm{x}_n]$, $W_s = \mathrm{diag}[w_1, \ldots, w_n]$ is the diagonal matrix with the sampling weights on the main diagonal and $T_s = \mathrm{diag}[t_1, \ldots, t_n]$, with the t_i 's defined by the estimators. For $\hat{\beta}_{\mathrm{ols}}$ $t_i = 1/w_i$, for $\hat{\beta}_{\mathrm{sel}}$ $t_i = w_i^{-1}\hat{\tau}_i^{-1}$ and so forth. The randomization variance of these estimators is estimated as,

$$\hat{\mathrm{Var}}_r(\hat{\beta}_t) = [X_s' W_s T_s X_s]^{-1} [\hat{\mathrm{Var}}_r \sum_{i=1}^{n} w_i t_i \mathrm{x}_i e_{it}] [X_s' W_s T_s X_s]^{-1}, \quad (4.3)$$

where $e_{it} = (y_i - \mathrm{x}_i' B)$ and B is the census estimator. Using the double index (hj) to define the j^{th} unit in the sample s_h of size n_h drawn from straum h, we estimated

$$\hat{\mathrm{Var}}_r\left[\sum_{i=1}^{n} w_i t_i \mathrm{x}_i e_{it}\right] = \sum_{h=1}^{5} \hat{\mathrm{Var}}\left(\sum_{j=1}^{n_h} w_{hj} \tilde{e}_{hj,t}\right)$$
$$= \sum_{h=1}^{5} \frac{n_h}{(n_h - 1)} \sum_{j=1}^{n_h} (w_{hj}\tilde{e}_{hj,t} - \overline{e}_{h,t})(w_{hj}\tilde{e}_{hj,t} - \overline{e}_{h,t})', \quad (4.4)$$

where $\tilde{e}_{hj,t} = t_{hj} \mathrm{x}_{hj} (y_{hj} - \mathrm{x}_{hj}' \hat{\beta}_t)$ and

$$\overline{e}_{h,t} = \frac{1}{n_h} \sum_{j=1}^{n_h} w_{hj} \tilde{e}_{hj,t},$$

assuming with replacement sampling within the strata.

A variance estimator under the sample model which accounts for possible heteroscedasticity is obtained as,

$$\hat{\mathrm{Var}}_{\mathrm{sm}}(\hat{\beta}_t) = [X_s' W_s T_s X_s]^{-1} \left[\sum_{i \in s} w_i^2 t_i^2 \hat{e}_{it}^2 \mathrm{x}_i \mathrm{x}_i'\right] [X_s' W_s T_s X_s]^{-1}, \quad (4.5)$$

where $\hat{e}_{it} = (y_i - \mathrm{x}_i' \hat{\beta}_t)$. Randomization and sample model variance estimators for the estimator in 4.1.2 are developed by Feder (2011). For the maximum likelihood estimator under the sample model with the likelihood defined by (4.1) we only estimate the variance under the sample model using the inverse information matrix.

Finally, bootstrap variance estimators for all the estimators are obtained by sampling with replacement n units from the original sample and re-estimating each of the estimators using the same computations as for the original sample. Repeating the same process independently B times, the bootstrap variance estimator is,

$$\hat{\mathrm{Var}}_{BS}(\hat{\beta}) = \frac{1}{B} \sum_{b=1}^{B} (\hat{\beta}^{(b)} - \overline{\hat{\beta}})(\hat{\beta}^{(b)} - \overline{\hat{\beta}})';$$
$$\overline{\hat{\beta}} = \frac{1}{B} \sum_{b=1}^{B} \hat{\beta}^{(b)}, \quad (4.6)$$

where $\hat{\beta}$ represents any of the estimators defined by 4.1.1 – 4.1.8 and $\hat{\beta}^{(b)}$ is the corresponding estimator computed for bootstrap sample b, $b = 1, \ldots, B$.

4.3 Computation of confidence intervals

We consider two approaches of $(1-\alpha)$ level confidence interval (C.I.) computation. The first approach is the standard C.I.,

$$\hat{\beta}_k \pm Z_{1-\frac{\alpha}{2}}\, \hat{s}.e(\hat{\beta}_k),\ k = 0, 1,$$

where $\hat{\beta}_k$ stands for any of the estimators considered and $\hat{s}.e(\hat{\beta}_k)$ is the corresponding estimator of the standard error as obtained by one of the methods listed before. The second, "basic bootstrap" approach uses the quantiles $bs(k, \tilde{\alpha})$ of the bootstrap estimators $\hat{\beta}_k^{(b)}$ to compute the C.I.

$$\left[2\hat{\beta}_k - bs\left(k, 1-\frac{\alpha}{2}\right), 2\hat{\beta}_k - bs\left(k, \frac{\alpha}{2}\right)\right], k = 1, 2.$$

We tried also the use of the "studentized bootstrap method" but the coverage rates were not better with any of the estimators $\hat{\beta}_k$. See Remark 14 below.

4.4 Simulation results

Table 1 shows the empirical means of the estimates listed in Section 4.1 over the 2,000 populations and samples and the corresponding empirical standard errors (S.E.). Also shown are the square roots of the means of the variance estimates as obtained when estimating the randomization variance ("Ran.") and when estimating the variance under the sample model ("S.M."). Because of computing time limitations, the results for the bootstrap variance estimators ("BS") are based on 300 bootstrap samples drawn from each of 500 original samples. These numbers of original and bootstrap samples were found to produce stable variance estimators.

As expected, given the use of an informative sampling scheme, the OLS estimator has a relatively large bias of 12% (5%) when estimating the intercept (slope). All the other estimators are virtually unbiased, except for $\hat{\beta}_{\text{mle}}$, which has bias of 2% and 1.5%. The almost unbiasedness of the EL estimator $\hat{\beta}_{\text{sel}}$ is particularly encouraging given the somewhat crude nonparametric estimation of the probabilities $\tau_i = \Pr(i \in s \mid y_i, \mathrm{x}_i)$. Notice also that this estimator has similar empirical S.E. to those of the PW estimator. The small (but statistically significant) bias of $\hat{\beta}_{\text{mle}}$ is explained by the fact that we assume a normal distribution under the population model, which as noted and illustrated before is incorrect.

Regarding precision, the OLS estimator has the smallest S.E. but $\hat{\beta}_f$ has almost the same S.E. (and is unbiased). This is explained by the fact that this estimator uses additional stratification information, not used by the other estimators. Note that $\hat{\beta}_{\text{mg}}, \hat{\beta}_{\text{mg-}q}$ and particularly $\hat{\beta}_q$ outperform $\hat{\beta}_{\text{pw}}$, but $\hat{\beta}_{\text{mg-}q}$ does not improve over $\hat{\beta}_q$.

Remark 14. Following my presentation of this paper at the 2011 Statistics Canada symposium, Jean-Francois Beaumont suggested to replace the weights $\hat{\tau}_i^{-1}$ used for the computation of $\hat{\beta}_{\text{sel}}$ by the weights $\hat{\tau}_i^{-1} / E_s(\hat{\tau}_i^{-1})$, so as to account for the net sampling effects on the conditional *pdf* $f(y \mid \mathrm{x})$, similarly to the use of the q-weights in $\hat{\beta}_q$. Notice that whereas the sampling weights w_i may depend on y, x and possibly other variables, the weights $\hat{\tau}_i^{-1}$ only depend on y and x. Application of this idea did not affect the bias but the empirical S.E. of the modified estimators are 0.151 and 0.053, smaller than the S.E. of $\hat{\beta}_{\text{sel}}$ and similar to the S.E. of $\hat{\beta}_q$.

Looking at the performance of the variance estimators, the first remarkable outcome is that the randomization and sample model variance estimators (Equations 4.4 and 4.5) are very similar for every estimator of the regression coefficients, even though they are computed very differently. For $\hat{\beta}_{\text{ols}}, \hat{\beta}_{\text{pw}}$ and $\hat{\beta}_q$ the variance estimators are almost unbiased but for the other estimators the variance estimators under-estimate the true variance. This is explained by the fact that these variance estimators ignore some of the operations involved in the computation of the estimated regression coefficients. Thus, in the case of the estimators $\hat{\beta}_{\text{mg}}$ and $\hat{\beta}_{\text{mg-}q}$ the variance estimators do not account for the choice of the optimal weights $a_i(\alpha)$, in the case of $\hat{\beta}_f$ the variance estimator does not account for the random imputation of the vectors (y_i, x_i) for $i \in U - s$, and in the case of $\hat{\beta}_{\text{mle}}$ and $\hat{\beta}_{\text{sel}}$ the variance estimators do not account for the estimation of the probabilities $\Pr(i \in s \mid y_i, \mathrm{x}_i)$. This under- estimation of the variance is corrected in almost all cases by use of the bootstrap method, see, in particular, the estimation of the variances of $\hat{\beta}_f, \hat{\beta}_{\text{mle}}$ and $\hat{\beta}_{\text{sel}}$.

Figure 2 shows the empirical coverage rates of $(1 - \alpha)$-level confidence intervals (C.I.) for $\alpha = 0.10, 0.05, 0.01$, as obtained when applying the standard C.I. with the standard errors estimated by the BS method, and when using the basic bootstrap method. The figures in the horizontal axis are the nominal levels

The coverage rates are almost always below the nominal levels but the under- coverage in the case of the standard C.I. is generally less than 4%. The two exceptions are when basing the confidence intervals on the OLS estimators (large under-coverage) and the mle estimator of the slope (under-coverage of 7% at the 90% nominal level), which is explained by the bias of these estimators. The under-coverage percentages when using the basic bootstrap method are generally slightly larger, except for the under-coverage of the C.I. for the intercept based on $\hat{\beta}_{\text{sel}}$, which is more pronounced.

Table 1
Means, standard errors (S.E.) and square roots of means of variance estimates. Population model: $E_p(y_j) = 2 + 1 \times x_j$, $\mathrm{Var}_p(y_j) = (1 + 0.2x_j)^2 \mathrm{V_j} + 1$

Method	Intercept- $\hat{\beta}_0$					Slope- $\hat{\beta}_1$				
	Mean Est.	Emp. S.E.	Ran.	S.M.	BS	Mean Est.	Emp. S.E.	Ran.	S.M.	BS
$\hat{\beta}_{\text{ols}}$	2.251	0.133	0.135	0.139	0.140	1.046	0.048	0.048	0.049	0.049
$\hat{\beta}_f$	2.006	0.133	0.126	0.126	0.135	0.999	0.051	0.041	0.041	0.052
$\hat{\beta}_{\text{pw}}$	2.008	0.166	0.167	0.169	0.157	0.998	0.059	0.055	0.055	0.056
$\hat{\beta}_{\text{mg}}$	2.017	0.158	0.154	0.156	0.154	0.995	0.056	0.050	0.050	0.055
$\hat{\beta}_q$	2.011	0.153	0.157	0.159	0.147	0.999	0.054	0.051	0.051	0.052
$\hat{\beta}_{\text{mg-}q}$	2.020	0.156	0.152	0.154	0.153	0.996	0.055	0.049	0.050	0.054
$\hat{\beta}_{\text{mle}}$	1.960	0.159	-----	0.143	0.152	1.026	0.054	-----	0.046	0.053
$\hat{\beta}_{\text{sel}}$	2.031	0.164	0.143	0.143	0.159	0.995	0.058	0.049	0.049	0.057

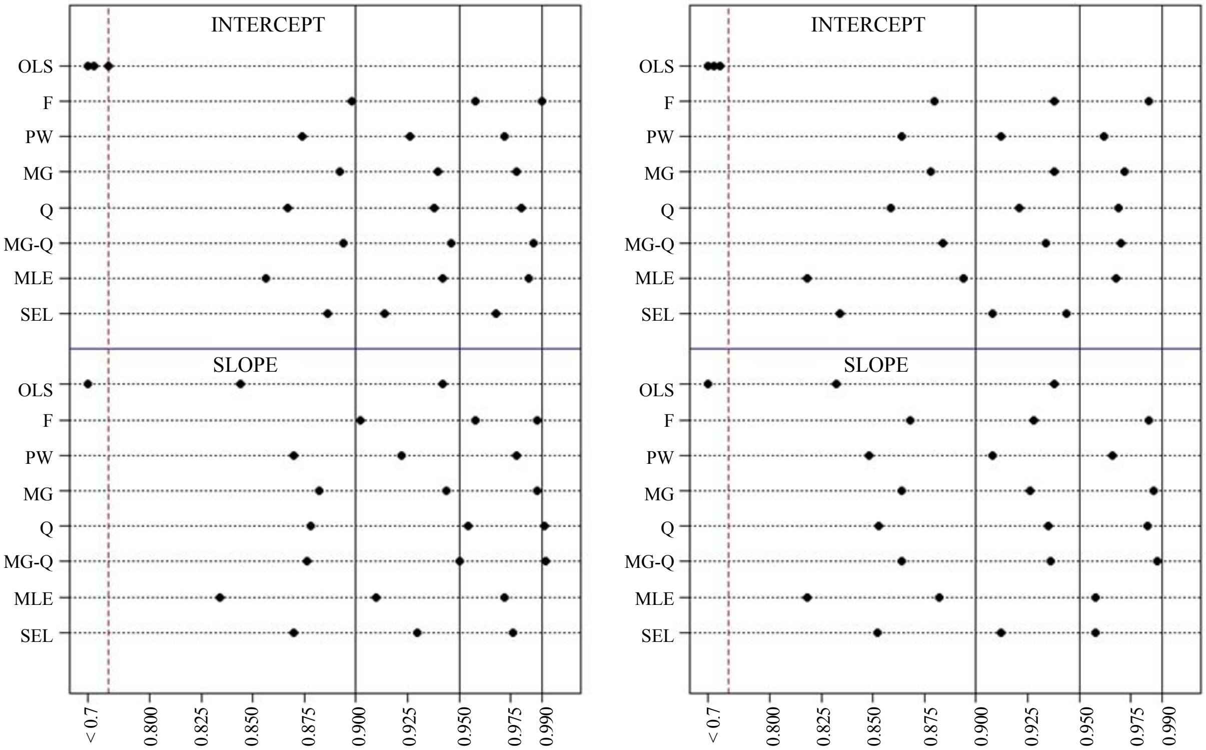


Figure 2 Coverage rates of standard (left) and BS (right) confidence intervals

Remark 15. We computed also the standard C.I. with the S.E. estimated under the randomization distribution (Equation 4.4) and under the sample model (Equation 4.5), but except in the case of the estimators $\hat{\beta}_{\text{pw}}$ and $\hat{\beta}_q$, the under-coverage of these intervals was somewhat higher than the coverage rates in Figure 2 because of the under-estimation of the true S.E. by these S.E. estimators discussed before. The same phenomenon was observed when using the "studentized bootstrap method" with these S.E. estimates, which again can be explained by the underestimation of the true S.E.'s. The use of more advanced bootstrap C.I. such as double-bootstrap may correct this under-coverage.

5. Concluding remarks

In this article I discuss alternative procedures proposed in the literature to account for informative sampling and NMAR nonresponse when modeling survey data. The empirical study is restricted so far to the case of linear regression and single-stage sampling, and an obvious extension would be to consider other models and cluster sampling. The present study illustrates the unbiasedness or approximate unbiasedness of all the point estimators considered, but the standard variance estimators underestimate the true variances in most cases since they fail to account for the extra operations involved in computing the corresponding point estimators. The bootstrap variance estimators produce much better variance estimators in these cases. The confidence intervals applied in the present study yield small under-coverage in most cases, but they should be improved, possibly by use of more advanced bootstrap techniques. Another important extension mentioned in the paper, which we have not investigated empirically so far is to incorporate sample based calibration constraints in the empirical likelihood method when based on the sample distribution.

We plan to apply the various methods to several real data sets. This would require the development of diagnostic procedures that would allow comparing the performance of the methods since unlike in a simulation study, the true distributions and model parameters are seldom known in real applications.

Acknowledgements

I am indebted to Dr. Moshe Feder for carrying out the empirical study and many helpful comments and suggestions. Thanks are due also to Dr. Pedro Silva for his constructive remarks on an earlier draft of the paper and three reviewers for their careful reading and comments in a short time period given to them. This study is funded by a UK ESRC grant No. RES-062-23-2316.

References

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

Binder, D., and Roberts, G. (2009). Design and model based inference for model parameters. In *Handbook of Statistics* 29B; *Sample Surveys*: *Inference and Analysis*, (Eds., D. Pfeffermann and C.R. Rao). Amsterdam: North Holland, 33-54.

Breckling, J.U., Chambers, R.L., Dorfman, A.H., Tam, S.M. and Welsh, A.H. (1994). Maximum likelihood inference from sample survey data. *International Statistical Review*, 62, 349-363.

Brick, J.M., and Montaquila, J.M. (2009). Nonresponse and weighting. In *Handbook of Statistics* 29A; *Sample Surveys*: *Inference and Analysis*, (Eds., D. Pfeffermann and C.R. Rao). Amsterdam: North Holland, 163-185.

Chambers, R.L., Dorfman, A.H. and Wang, S. (1998). Limited information likelihood analysis of survey data. *Journal of the Royal Statistical Society*, Series B, 60, 397-411.

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

Chambless, L.E., and Boyle, K.E. (1985). Maximum likelihood methods for complex sample data: Logistic, regression and discrete proportional hazards models. *Communication in Statistics-Theory and Methods*, 14, 1377-1392.

Chang, T., and Kott, P.S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika*, 95, 555-571.

Chen, J., and Sitter, R.R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica sinica*, 9, 385-406.

Chaudhuri, S., Handcock, M.S. and Rendall, M.S. (2010). A conditional empirical likelihood approach to combine sampling design and population level information. Technical report No. 3/2010, National University of Singapore, Singapore, 117546.

DeMets, D., and Halperin, M. (1977). Estimation of simple regression coefficients in samples arising from sub-sampling procedures. *Biometrics*, 33, 47-56.

DuMouchel, W.H., and Duncan, G.L. (1983). Using sample survey weights in multiple regression analysis of stratified samples. *Journal of the American Statistical Association*, 78, 535-543.

Feder, M. (2011). Fitting Regression Models to Complex Survey Data- Gelman's Estimator Revisited. In Proceedings of the ISI meeting, Ireland, (www.isi2011.ie).

Francisco, C.A., and Fuller, W.A. (1991). Quantile estimation with a complex survey design. *The Annals of Statistics*, 19, 454-469.

Fuller, W.A. (1975). Regression analysis for sample surveys. *Sankhyā*, Series C, 37, 117-132.

Fuller, W.A. (2002). Regression estimation for survey samples. *Survey Methodology*, 28, 5-23.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2003). *Bayesian Data Analysis*, second edition. London: CRC Press.

Gelman, A. (2007). Struggles with survey weighting and regression modeling (with discussion). *Statistical Science*, 22, 153-164.

Godambe, V.P., and Thompson, M.E. (1986). Parameters of superpopulation and survey population: Their relationships and estimation. *International Statistical Review*, 54, 127-138.

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

Goldstein, H. (1986). Multi-level mixed linear model analysis using iterative generalized least squares. *Biometrika*, 73, 43-56.

Hàjek, J. (1971). Comments on a paper by D. Basu. In *Foundations of Statistical Inference*, (Eds., V.P. Godambe and D.A. Sprott). Toronto: Holt, Rinehart and Winston.

Hartley, H.O., and Rao, J.N.K. (1968). A new estimation theory for sample surveys. *Biometrika*, 55, 547-557.

Holt, D., Smith, T.M.F. and Winter, P.D. (1980). Regression analysis of data from complex surveys. *Journal of the Royal Statistical Society*, Series A, 143, 474-487.

Jewell, N.P. (1985). Least squares regression with data arising from stratified samples of the dependent variable. *Biometrika*, 72, 11-21.

Kasprzyk, D., Duncan, G.J., Kalton, G. and Singh, M.P. (1989, Eds.). *Panel Surveys*. New York: John Wiley & Sons, Inc.

Kim, J.K. (2009). Calibration estimation using empirical likelihood in survey sampling. *Statistica Sinica*, 19, 145-157.

Kott, P.S. (2009). Calibration Weighting: Combining Probability Samples and Linear Prediction Models. In *Handbook of Statistics* 29B; *Sample Surveys*: *Inference and Analysis*, (Eds., D. Pfeffermann and C.R. Rao). Amsterdam: North Holland, 55-82.

Krieger, A.M., and Pfeffermann D. (1997). Testing of distribution functions from complex sample surveys. *Journal of Official Statistics*, 13, 123-142.

Little, R.J.A. (1982). Models for non-response in sample surveys. *Journal of the American Statistical Association*, 77, 237-249.

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

Magee, L. (1998). Improving survey-weighted least squares regression. *Journal of the Royal Statistical Society*, Series B, 60, 115-126.

Nathan, G., and Holt, D. (1980). The effect of survey design on regression analysis. *Journal of the Royal Statistical Society*, Series B, 42, 377-386.

Orchard, T., and Woodbury, M.A. (1972). A missing information principle: Theory and application. *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability*, 1, 697-715.

Owen, A.B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75, 237-249.

Owen, A.B. (2001). *Empirical likelihood*. New York: Chapman & Hall.

Pfeffermann, D., and Holmes, D. (1985). Robustness consideration in the choice of method of inference for regression analysis of survey data. *Journal of the Royal Statistical Society*, Series A, 148, 268-278.

Pfeffermann, D., and Smith, T.M.F. (1985). Regression models for grouped populations in cross-section surveys. *International Statistical Review*, 53, 37-59.

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

Pfeffermann, D. (1996). The use of sampling weights for survey data analysis. *Statistical Methods in Medical Research*, 5, 239-261.

Pfeffermann, D., Krieger, A.M. and Rinott, Y. (1998a). Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica*, 8, 1087-1114.

Pfeffermann, D., Skinner, C.J., Holmes, D.J., Goldstein, H. and Rasbash, J. (1998b). Weighting for unequal selection probabilities in multi-level models (with discussion). *Journal of the Royal Statistical Society*, Series B, 60, 23-76.

Pfeffermann, D., and Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhyā*, 61, 166-186.

Pfeffermann, D., and Sverchkov, M. (2003). Fitting generalized linear models under informative probability sampling. In *Analysis of Survey Data*, (Eds., R.L. Chambers and C.J. Skinner). New York: John Wiley & Sons, Inc., 175-195.

Pfeffermann, D., Moura, F.A.S. and Nascimento-Silva, P.L. (2006). Multilevel modeling under informative sampling. *Biometrika*, 93, 943-959.

Pfeffermann., D., and Sverchkov, M. (2007). Small area estimation under informative probability sampling of areas and within the selected areas. *Journal of the American Statistical Association*, 102, 1427-1439.

Pfeffermann, D., and Sverchkov, M. (2009). Inference under Informative Sampling. In *Handbook of Statistics* 29B; *Sample Surveys*: *Inference and Analysis*, (Eds., D. Pfeffermann and C.R. Rao). Amsterdam: North Holland, 455-487.

Pfeffermann, D., and Landsman, V. (2011). Are private schools better than public schools? Appraisal for Ireland by methods for observational studies. *The Annals of Applied Statistics*, 5, 1726-1751.

Pfeffermann, D., and Sikov, N. (2011). Imputation and estimation under nonignorable nonresponse in household surveys with missing covariate information. *Journal of Official Statistics*, 27, 181-209.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 605-614.

Rubin, D.B. (1985). The use of propensity scores in applied Bayesian inference. In *Bayesian Statistics* 2, (Eds., J.M. Bernardo, M.H. Degroot, D.V. Lindley and A.F.M. Smith), Elsevier Science Publishers B.V., 463-472.

Särndal, C.-E., and Wright, R. (1984). Cosmetic form of estimators in survey sampling. *Scandinavian Journal of Statistics*, 11, 146-156.

Scott, A.J., and Holt, D. (1982). The effect of two-stage sampling on ordinary least squares. *Journal of the American Statistical Association*, 77, 848-854.

Scott, A.J., and Wild, C.J. (2009). Population-based case-control studies. In *Handbook of Statistics* 29B; *Sample Surveys*: *Inference and Analysis*, (Eds., D. Pfeffermann and C.R. Rao). Amsterdam: North Holland, 431-453.

Skinner, C.J., Holt, D. and Smith, T.M.F. (Eds.) (1989). *Analysis of complex surveys*. New York: John Wiley & Sons, Inc.

Skinner, C.J. (1994). Sample models and weights. *Proceedings of the Section on Survey Research Methods*, 133-142.

Smith, T.M.F. (1988). To weight or not to weight, that is the question. In *Bayesian Statistics* 3, (Eds., J.M. Bernardo, M.H. Degroot, D.V. Lindley and A.F.M. Smith), Oxford University Press, 437-451.

Sugden, R.A., and Smith, T.M.F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika*, 71, 495-506.

Sverchkov, M., and Pfeffermann, D. (2004). Prediction of finite population totals based on the sample distribution. *Survey Methodology*, 30, 79-92.

Wu, Y.Y., and Fuller, W.A. (2006). Estimation of regression coefficients with unequal probability samples. *Proceedings of the American Statistical Association*, *Section on Survey Research Methods*, 3892-3899.

Survey Methodology, December 2011
Vol. 37, No. 2, pp. 137-152
Statistics Canada, Catalogue No. 12-001-X

A Bayesian analysis of small area probabilities under a constraint

Balgobin Nandram and Hasanjan Sayit [1]

Abstract

In many sample surveys there are items requesting binary response (*e.g.*, obese, not obese) from a number of small areas. Inference is required about the probability for a positive response (*e.g.*, obese) in each area, the probability being the same for all individuals in each area and different across areas. Because of the sparseness of the data within areas, direct estimators are not reliable, and there is a need to use data from other areas to improve inference for a specific area. Essentially, a priori the areas are assumed to be similar, and a hierarchical Bayesian model, the standard beta-binomial model, is a natural choice. The innovation is that a practitioner may have much-needed additional prior information about a linear combination of the probabilities. For example, a weighted average of the probabilities is a parameter, and information can be elicited about this parameter, thereby making the Bayesian paradigm appropriate. We have modified the standard beta-binomial model for small areas to incorporate the prior information on the linear combination of the probabilities, which we call a constraint. Thus, there are three cases. The practitioner (a) does not specify a constraint, (b) specifies a constraint and the parameter completely, and (c) specifies a constraint and information which can be used to construct a prior distribution for the parameter. The griddy Gibbs sampler is used to fit the models. To illustrate our method, we use an example on obesity of children in the National Health and Nutrition Examination Survey in which the small areas are formed by crossing school (middle, high), ethnicity (white, black, Mexican) and gender (male, female). We use a simulation study to assess some of the statistical features of our method. We have shown that the gain in precision beyond (a) is in the order with (b) larger than (c).

Key Words: Accept-reject algorithm; Binomial distribution; Generalized beta distribution; Griddy Gibbs sampler; Simulation.

1. Introduction

It is a standard practice to use models to “borrow strength” in small area estimation (Rao 2003). Owing to the sparseness of the data in each area, direct estimates for small areas are typically not reliable. Our procedure allows a practitioner to incorporate prior information about a linear combination of binomial probabilities, one for each area. This is a constraint that we include as a weighted average of the area probabilities in the standard beta-binomial model. The weighted average can be assumed known or unknown. In the case when this value is unknown, we consider the scenario when there is some information which can be elicited from an expert in the form of prior distribution. This is different from standard practice in design based survey sampling in which auxiliary information is incorporated as in ratio and regression estimators (Cochran 1977). When the value can be specified exactly, there will be an increase in precision because prior information is incorporated into the model.

The beta-binomial model has been studied extensively. For example, Nandram and Sedransk (1993), Nandram (1998) and Nandram and Choi (2002) show how to do Bayesian predictive inference of finite population proportions of the small areas for binomial and multinomial data. These models assume that the binomial probabilities share a common effect, thereby permitting adaptive pooling of the data from small areas (or clusters). However, it is possible to improve on these models further by including additional information using covariates via generalized linear models (*e.g.*, see Ghosh, Natarajan, Stroud and Carlin 1998). It is worth noting that none of these works propose ways to incorporate prior information about linear combination of model parameters. Substantial gains in precision are expected when such prior information is incorporated in small area models; see Silvapulle and Sen (2006) for a book-length discussion of constrained statistical inference. It is also worth noting that Lazar, Meeden and Nelson (2008) showed how to include constraints in nonparametric Bayesian approach via a Polya urn scheme to predictive distribution of finite population parameters.

Our procedure is related to external benchmarking which occurs when a pre-specified estimator is obtained from external sources, such as a different survey, a census, or other administrative records. In benchmarking one wants the parts to add up to the whole. For example, when surveys are conducted over time, there are typically monthly surveys and annual surveys which are of much better quality than the monthly surveys. When the monthly surveys are estimated such that these estimates add up to the annual survey totals, there is a protection against model failure and therefore improved estimates (*i.e.*, reduced bias and possibly an increase in precision). These problems are prevalent in the government agencies especially in employment and sales;

[1] Balgobin Nandram and Hasanjan Sayit, Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609-2280. E-mail: balnan@wpi.edu, hs7@wpi.edu.

see Hillmer and Trabelsi (1987) for an example on retail sales of hardware stores from the U.S. Census Bureau.

Prior information from external benchmarking will lead to improved precision but can produce severely biased estimators as well. This will depend on how different the current survey is from the prior ones. Nandram, Toto and Choi (2011) applied external benchmarking to estimate the finite population means of small areas. The constraint is the finite population mean for the entire population is a prespecified value which again can be obtained from a prior survey, census or administrative records. In our current work we are not incorporating information about a linear combination of the finite population values, but rather we are inputting information about a linear combination of the superpopulation parameters (in this case binomial probabilities).

We consider the problem in which binomial counts are obtained from similar small areas, and inference is required about the binomial probabilities. In the conclusion, we discuss how to extend our method to obtain the predictive distribution of finite population proportions. The standard beta-binomial model may be inadequate, and additional prior information must be incorporated. Our thesis is that there is an increase in precision over the standard beta-binomial small area model when prior information about the weighted average of the probabilities (*e.g.*, average of the probabilities) is incorporated. That is, we incorporate prior information about a linear combination of binomial probabilities (a weighted average). The weights can be proportional to population sizes, and under proportional allocation they can be proportional to the sample sizes themselves. The purpose of incorporating prior information about the binomial probabilities is to increase precision, and at the same time one needs to control the bias.

It is much easier for a survey practitioner to specify the value of the overall probability rather than the individual area probabilities. That is, the overall probability can be specified with relatively much less error than the individual probabilities. Of course, one can specify the overall probability using prior information (a prior survey, census or administrative records), and so the specification of the overall probability will depend on the quality of the prior information. Thus, the problem falls naturally within the Bayesian paradigm because we are incorporating prior information about a parameter via a distribution. Thus, there will be gains in precision because of the extra information. However, a practitioner can still proceed when there is no prior information. One can use the ratio of the total success and total sample size over areas to form a reasonable specification of the overall probability which is typically not of interest. This estimate will have much higher precision than the one for individual areas. There will still be gain in precision, but clearly such gain is due to using the current data (double use) and the constraint.

One example of a survey in which reliable information can be obtained to perform the benchmarking is the National Health Interview Survey (NHIS) which is conducted annually by the National Center for Health Statistics to assess an aspect of Health of the U.S. population. This is a population-based survey and there are many health indicators of interest; one of these indicators is the number of doctor visits made in the past two weeks, and an informative quantity is the proportion of people who made at least one doctor visit last year (*e.g.*, Nandram and Choi 2002). These proportions are useful for small domains formed by crossing age, race and sex for a particular state last year. Because the estimates over a state change very slowly over the previous years, the overall estimate from the year immediately preceding last year can be used as a reliable benchmark for last year. If a reliable estimate cannot be obtained for the benchmark, one can construct an informative prior distribution for it. For example, one can use the method of moments to equate the sample mean and sample variance of the overall estimates for the past few years to the mean and variance of a beta distribution to get a beta prior distribution. In either case, our procedure can be applied.

The plan of this paper is as follows. In Section 2 we describe the methodology. Specifically, we describe the standard beta-binomial model, and we develop two additional models to incorporate the extra information using appropriate prior distributions. We also describe posterior inference and how to perform the nonstandard computations. In Section 3 we describe an illustrative example on obesity, and a simulation study to assess empirically the statistical properties of our models. Section 4 has concluding remarks. We also discuss how to do Bayesian predictive inference for finite population proportions. While we discuss binary data, we also show how one can extend our method to polychotomous data.

2. Methodology

We show how to incorporate the constraint into the beta-binomial model in two ways, thereby providing a set of alternative models. In Section 2.1 we describe the models and in Section 2.2 we describe posterior inference. We attempt to explain what the constraint does to the estimates of the probabilities using an approximation. In Section 2.3 we describe the computation, and we describe a new algorithm as well.

2.1 Models

We assume that binary data are available from ℓ small areas, and we assume that the probability that an individual responds in the i^{th} area is π_i, $i = 1, ..., \ell$. Let n_i be the number of individuals sampled from the i^{th} area, $i = 1, ..., \ell$. Also let s_i denote the number of individuals with the characteristic and $f_i = n_i - s_i$ be the number of individuals without the characteristic in the i^{th} area, $i = 1, 2, ..., \ell$. Then the standard beta-binomial hierarchical Bayesian model is

$$s_i \mid \pi_i \overset{\text{ind}}{\sim} \text{Binomial}\,(n_i, \pi_i), \tag{1}$$

$$\pi_i \mid \mu, \tau \overset{\text{iid}}{\sim} \text{Beta}\,\{\mu\tau, (1-\mu)\tau\},\; i = 1, ..., \ell \tag{2}$$

and

$$p(\mu, \tau) = \frac{1}{(1+\tau)^2},\; 0 < \mu < 1,\; \tau \ge 0. \tag{3}$$

We use a shrinkage prior for τ because it is proper and noninformative, and there are no conjugate priors. Priors of the form $p(\tau) \propto 1/\tau$ are discouraged; see, for example, Gelman (2006). Other alternatives are half Cauchy densities and gamma densities (one would need to specify the hyperparameters). Henceforth, we will call the model specified by (1), (2) and (3) the unrestricted (UR) model or Model 1.

We next describe the restricted model, which is an extension of the unrestricted model. We obtain a simple linear combination of the binomial probabilities. Letting $\tilde{\pi}_i = s_i/n_i$ and

$$\omega_i = \frac{n_i}{\sum_i^{\ell} n_i},\; i = 1, ..., \ell,$$

we have

$$\frac{\sum_{i=1}^{\ell} s_i}{\sum_{i=1}^{\ell} n_i} = \sum_{i=1}^{\ell} \omega_i\, \tilde{\pi}_i.$$

Thus, taking the π_i unknown, the linear combination is $\sum_{i=1}^{\ell} \omega_i\, \pi_i$.

Therefore, we need to make an adjustment in (2) to incorporate the restriction, $\sum_{i=1}^{\ell} \omega_i \pi_i = \theta$ conditional on θ. We do so by introducing the variable $\phi = \sum_{i=1}^{\ell} \omega_i\, \pi_i - \theta$; so that the restriction is equivalent to $\phi = 0$. Now one of the variables, $\pi_i, i = 1, \ldots, \ell$, is redundant. It is worth noting that one can choose any one of $\pi_1, ..., \pi_\ell$, and without loss of generality and for ease of exposition, we choose π_ℓ. Thus, to incorporate the restriction, we transform π_ℓ to $\phi = \sum_{i=1}^{\ell} \omega_i\, \pi_i - \theta$, keeping $\pi_1, ..., \pi_{\ell-1}$ untransformed, and we let $\boldsymbol{\pi}_{(\ell)} = (\pi_1, \ldots, \pi_{\ell-1})'$.

As the jacobian is $1/\omega_\ell$,

$$p(\boldsymbol{\pi}_{(\ell)}, \phi \mid \mu, \tau, \theta) =$$

$$\frac{1}{\omega_\ell} \prod_{i=1}^{\ell-1} \frac{\pi_i^{\mu\tau-1}(1-\pi_i)^{(1-\mu)\tau-1}}{B\{\mu\tau, (1-\mu)\tau\}}$$

$$\times \frac{\left[\dfrac{\phi + \theta - \sum_{i=1}^{\ell-1}\omega_i \pi_i}{\omega_\ell}\right]^{\mu\tau-1} \left[1 - \dfrac{\phi + \theta - \sum_{i=1}^{\ell-1}\omega_i \pi_i}{\omega_\ell}\right]^{(1-\mu)\tau-1}}{B\{\mu\tau, (1-\mu)\tau\}}, \tag{4}$$

where

$$0 < \pi_i < 1,\; i = 1, ..., \ell,$$

$$0 < \mu < 1,\; \tau > 0,\; \phi + \theta - \omega_\ell \le \sum_{i=1}^{\ell-1} \omega_i\, \pi_i \le \phi + \theta,$$

and

$$\pi_\ell = \frac{\phi + \theta - \sum_{i=1}^{\ell-1} \omega_i\, \pi_i}{\omega_\ell}. \tag{5}$$

Note that the joint prior density of $(\boldsymbol{\pi}_{(\ell)}, \phi)$ in (4) is well defined. We wish to take $\phi = 0$ in (5) to incorporate the restriction, but when $\phi = 0$ the joint density of $\boldsymbol{\pi}_{(\ell)}$ is not well defined.

We assume μ, τ, θ are independent a priori with $p(\mu, \tau, \theta) = p_1(\mu, \tau)\, p_2(\theta)$, where

$$p_1(\mu, \tau) = \frac{1}{(1+\tau)^2},\quad 0 < \mu < 1,\; \tau \ge 0$$

as in (3), and $p_2(\theta)$ is given by

$$\theta \sim \text{Beta}\,\{\mu_0\tau_0, (1-\mu_0)\tau_0\}. \tag{6}$$

For the restricted model we consider two scenarios. Letting $\tau_0 \to \infty$, θ becomes a point mass at μ_0, and in this case $\theta = \mu_0$ is to be specified by a practitioner; we will call the adjusted model the fixed (FI) model or Model 2. We have a second scenario in which a practitioner specifies μ_0 and τ_0 but not θ; we will call this adjusted model the informative (IN) model or Model 3. Thus, there are three models, including the unrestricted model. To provide a unified framework, we need all our priors to be proper. The exact value of θ is likely to be unknown in most applications, and this can lead to estimates which are not internally coherent.

It is worth noting that we have considered an additional model to help study the gain in precision of IN relative to FI. For comparison we want to impose a proper but noninformative prior on θ, so that $\theta \sim \text{Uniform}\,(0, 1)$ is not an unreasonable choice. Letting $\mu_0 = 1/2, \tau_0 = 2$, we get $\theta \sim \text{Uniform}\,(0, 1)$ with this prior, and we will call the adjusted model the uniform (UN) model or Model 4; of course, we do not need to specify μ_0 and τ_0. It is worth noting that the prior corresponding to $\tau \to \infty$ is improper as it corresponds to $\theta \sim$ Beta (0, 0). We do not consider this model further; however, although UN does not have a constraint, we will consider it briefly throughout.

2.2 Posterior inference

We consider making posterior inference about π_i, $i = 1, \ldots, \ell$. Let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_\ell)'$ and $\boldsymbol{\pi}_{(i)} = (\pi_1, \ldots, \pi_{i-1}, \pi_{i+1}, \ldots, \pi_\ell)'$ [*e.g.*, $\boldsymbol{\pi}_{(\ell)} = (\pi_1, \ldots, \pi_{\ell-1})'$ as defined above].

We use Bayes' theorem to find the joint posterior densities of all parameters. First, under the unrestricted model specified by (1), (2) and (3) the joint posterior density of $\boldsymbol{\pi}, \mu, \tau$ is

$$g(\boldsymbol{\pi}, \mu, \tau \mid \boldsymbol{s}) \propto \prod_{i=1}^{\ell} \frac{\pi_i^{s_i+\mu\tau-1}(1-\pi_i)^{f_i+(1-\mu)\tau-1}}{B\{s_i+\mu\tau,\ f_i+(1-\mu)\tau\}}$$

$$\times \prod_{i=1}^{\ell} \frac{B\{s_i+\mu\tau,\ f_i+(1-\mu)\tau)\}}{B\{\mu\tau, (1-\mu)\tau)\}}$$

$$\times \frac{1}{(1+\tau)^2}, \tag{7}$$

$0 < \pi_i < 1,\ 0 < \mu < 1,\ \tau > 0,\ i = 1, \ldots, \ell.$

Lemma 1 *Under the unrestricted model the joint posterior density,* $g(\boldsymbol{\pi}, \mu, \tau \mid \boldsymbol{s})$, *is proper.*

A proof of Lemma 1 is given in Appendix A.

Under the restricted model the joint posterior density of $\boldsymbol{\pi}_{(\ell)}, \mu, \tau, \theta, \phi$ is

$p(\boldsymbol{\pi}_{(\ell)}, \mu, \tau, \theta, \phi \mid \boldsymbol{s})$

$$\propto \prod_{i=1}^{\ell-1} \frac{\pi_i^{s_i+\mu\tau-1}(1-\pi_i)^{f_i+(1-\mu)\tau-1}}{B\{s_i+\mu\tau,\ f_i+(1-\mu)\tau\}}$$

$$\times \frac{\left[\dfrac{\phi+\theta-\sum_{i=1}^{\ell-1}\omega_i\pi_i}{\omega_\ell}\right]^{s_\ell+\mu\tau-1}\left[1-\dfrac{\phi+\theta-\sum_{i=1}^{\ell-1}\omega_i\pi_i}{\omega_\ell}\right]^{f_\ell+(1-\mu)\tau-1}}{B\{s_\ell+\mu\tau,\ f_\ell+(1-\mu)\tau\}}$$

$$\times \prod_{i=1}^{\ell}\left[\frac{B\{s_i+\mu\tau,\ f_i+(1-\mu)\tau\}}{B\{\mu\tau, (1-\mu)\tau\}}\right]$$

$$\times \theta^{\mu_0\tau_0-1}(1-\theta)^{(1-\mu_0)\tau_0-1} \times \frac{1}{(1+\tau)^2}, \tag{8}$$

$0 < \pi_i < 1,\ i = 1, \ldots, \ell,\ 0 < \mu < 1,\ \tau > 0,\ \phi + \theta - \omega_\ell \le \sum_{i=1}^{\ell-1}\omega_i\pi_i \le \phi + \theta,\ 0 < \theta < 1.$ Note that $\pi_\ell = (\phi + \theta - \sum_{i=1}^{\ell-1}\omega_i\pi_i) / \omega_\ell$.

We get the pertinent joint posterior density by incorporating the constraint ($\phi = 0$) into (8). That is, $p(\boldsymbol{\pi}_{(\ell)}, \mu, \tau, \theta \mid \boldsymbol{s}, \phi = 0) \propto p(\boldsymbol{\pi}_{(\ell)}, \mu, \tau, \theta, \phi = 0 \mid \boldsymbol{s})$, where

$p(\boldsymbol{\pi}_{(\ell)}, \mu, \tau, \theta \mid \boldsymbol{s}, \phi = 0)$

$$\propto \prod_{i=1}^{\ell-1} \frac{\pi_i^{s_i+\mu\tau-1}(1-\pi_i)^{f_i+(1-\mu)\tau-1}}{B\{s_i+\mu\tau,\ f_i+(1-\mu)\tau\}}$$

$$\times \frac{\left[\dfrac{\theta-\sum_{i=1}^{\ell-1}\omega_i\pi_i}{\omega_\ell}\right]^{s_\ell+\mu\tau-1}\left[1-\dfrac{\theta-\sum_{i=1}^{\ell-1}\omega_i\pi_i}{\omega_\ell}\right]^{f_\ell+(1-\mu)\tau-1}}{B\{s_\ell+\mu\tau,\ f_\ell+(1-\mu)\tau\}}$$

$$\times \prod_{i=1}^{\ell}\left[\frac{B\{s_i+\mu\tau,\ f_i+(1-\mu)\tau\}}{B\{\mu\tau, (1-\mu)\tau\}}\right]$$

$$\times \theta^{\mu_0\tau_0-1}(1-\theta)^{(1-\mu_0)\tau_0-1} \times \frac{1}{(1+\tau)^2}, \tag{9}$$

$0 < \pi_i < 1,\ i = 1, \ldots, \ell,\ 0 < \mu < 1,\ \tau > 0,\ \theta - \omega_\ell \le \sum_{i=1}^{\ell-1}\omega_i\,\pi_i \le \theta,\ 0 < \theta < 1.$ Note again that $\pi_\ell = (\theta - \sum_{i=1}^{\ell-1}\omega_i\pi_i) / \omega_\ell$. It is worth noting that the joint posterior density (9) incorporates the constraint, $\sum_{i=1}^{\ell}\omega_i\pi_i = \theta$, exactly because $\pi_\ell = (\theta - \sum_{i=1}^{\ell-1}\omega_i\pi_i) / \omega_\ell$, $\theta - \omega_\ell \le \sum_{i=1}^{\ell-1}\omega_i\pi_i \le \theta$. That is, the joint posterior density is not a function of π_ℓ, and posterior inference about π_ℓ follows from the identity, $\pi_\ell = (\theta - \sum_{i=1}^{\ell-1}\omega_i\pi_i) / \omega_\ell$. Thus, there is absolutely no difference between θ and $\sum_{i=1}^{\ell}\omega_i\pi_i$.

Theorem 1 *Under the restricted model the joint posterior density,* $p(\boldsymbol{\pi}_{(\ell)}, \mu, \tau, \theta \mid \boldsymbol{s}, \phi = 0)$, *is proper.*

A proof of Theorem 1 is given in Appendix A.

We note the difference between the densities for the unrestricted model in (7) and the restricted model in (9). Essentially, the term

$$\left(\frac{\theta-\sum_{i=1}^{\ell-1}\omega_i\pi_i}{\omega_\ell}\right)^{s_\ell+\mu\tau-1} \times \left(1-\frac{\theta-\sum_{i=1}^{\ell-1}\omega_i\pi_i}{\omega_\ell}\right)^{f_\ell+(1-\mu)\tau-1}$$

$$\times\, \theta^{\mu_0\tau_0-1}(1-\theta)^{(1-\mu_0)\tau_0-1}$$

in (9) replaces $\pi_\ell^{s_\ell+\mu\tau-1}(1-\pi_\ell)^{f_\ell+(1-\mu)\tau-1}$ in (7). Note that in (9),

$$\pi_\ell = \frac{\theta - \sum_{i=1}^{\ell-1}\omega_i\pi_i}{\omega_\ell}.$$

Let $a_i = s_i + \mu\tau$, $b_i = f_i + (1-\mu)\tau$, $i = 1, \ldots, \ell$. Also let

$$c_i = \frac{\theta - \sum_{j=1, j\ne i}^{\ell-1}\omega_j\pi_j - \omega_\ell}{\omega_i}$$

and

$$d_i = \frac{\theta - \sum_{j=1, j \neq i}^{\ell-1} \omega_j \pi_j}{\omega_i}, \; i = 1, ..., \ell - 1.$$

Then,

$$\begin{aligned} &p(\pi_i \mid \boldsymbol{\pi}_{(i)}, \mu, \tau, \theta, \boldsymbol{s}, \phi = 0) \\ &\qquad \propto \pi_i^{a_i - 1}(1 - \pi_i)^{b_i - 1}(\pi_i - c_i)^{b_\ell - 1}(d_i - \pi_i)^{a_\ell - 1}, \qquad (10) \end{aligned}$$

$c_i < \pi_i < d_i, i = 1, ..., \ell - 1.$ Note that this density function consists of two terms $\pi_i^{a_i - 1}(1 - \pi_i)^{b_i - 1}$ and $(\pi_i - c_i)^{b_\ell - 1}$ $(d_i - \pi_i)^{a_\ell - 1}$; note the interchange between a_ℓ and b_ℓ in the second term. The first term is the conditional posterior density under the unrestricted model, and the second term is a generalized beta density [*i.e.*, a $\text{beta}(b_\ell, a_\ell)$ distribution in the interval (c_i, d_i)]. Thus, the unrestricted beta density is adjusted by the generalized beta density. In the rest of the paper we denote by $\text{GenBeta}(a, b, c, d)$ the generalized beta random variable with density function,

$$p(x) = (x - c)^{a-1}(d - x)^{b-1} / \{(d - c)^{a+b-1} B(a, b)\},$$
$$c \leq x \leq d, a > 1, b > 1.$$

That is, $(X - c) / (d - c) \sim \text{Beta}(a, b)$ if and only if $X \sim \text{GenBeta}(a, b, c, d)$.

It is worth noting that we have ordered the areas in order of their counts (smallest to largest). This is convenient and advantageous both theoretically and computationally.

In order to explain the gain in precision, we attempt to study (10) further by making two approximations. First, because the restriction under study is rather mild we do not expect c_i to be much different from 0 and d_i to be much different from 1. Under this assumption, we can approximate (10) by

$$\begin{aligned} &p_a(\pi_i \mid \boldsymbol{\pi}_{(i)}, \mu, \tau, \theta, \boldsymbol{s}, \phi = 0) \\ &\qquad \propto (\pi_i - c_i)^{a_i - 1}(d_i - \pi_i)^{b_i - 1}(\pi_i - c_i)^{b_\ell - 1}(d_i - \pi_i)^{a_\ell - 1}, \\ &\qquad c_i < \pi_i < d_i. \end{aligned}$$

Then, incorporating the normalization constant into $p_a(\pi_i \mid \boldsymbol{\pi}_{(i)}, \mu, \tau, \theta, \boldsymbol{s}, \phi = 0)$, we have

$$\begin{aligned} &p_a(\pi_i \mid \boldsymbol{\pi}_{(i)}, \mu, \tau, \theta, \boldsymbol{s}, \phi = 0) \\ &= \frac{(\pi_i - c_i)^{a_i - 1}(d_i - \pi_i)^{b_i - 1}(\pi_i - c_i)^{b_\ell - 1}(d_i - \pi_i)^{a_\ell - 1}}{\int_{c_i}^{d_i} (\pi_i - c_i)^{a_i - 1}(d_i - \pi_i)^{b_i - 1}(\pi_i - c_i)^{b_\ell - 1}(d_i - \pi_i)^{a_\ell - 1} d\pi_i} \\ &= \frac{(\pi_i - c_i)^{a_i - 1}(d_i - \pi_i)^{b_i - 1}}{(d_i - c_i)^{a_i + b_i - 1} B(a_i, b_i)} \\ &\times \frac{(\pi_i - c_i)^{b_\ell - 1}(d_i - \pi_i)^{a_\ell - 1}}{E[(\pi_i - c_i)^{b_\ell - 1}(d_i - \pi_i)^{a_\ell - 1}]}, \; c_i < \pi_i < d_i, \qquad (11) \end{aligned}$$

where the expectation is taken over the generalized Beta distribution $\pi_i \sim \text{GenBeta}(a_i, b_i, c_i, d_i), i = 1, ..., \ell - 1.$ But under this latter density, $(\pi_i - c_i)^{b_\ell - 1}(d_i - \pi_i)^{a_\ell - 1}$ is an unbiased estimator of $E[(\pi_i - c_i)^{b_\ell - 1}(d_i - \pi_i)^{a_\ell - 1}].$ In addition, by construction a_ℓ and b_ℓ are relatively large and therefore $(\pi_i - c_i)^{b_\ell - 1}(d_i - \pi_i)^{a_\ell - 1}$ and its variance are expected to be small. Then, our second approximation is

$$(\pi_i - c_i)^{b_\ell - 1}(d_i - \pi_i)^{a_\ell - 1} \approx E[(\pi_i - c_i)^{b_\ell - 1}(d_i - \pi_i)^{a_\ell - 1}]. \quad (12)$$

Therefore, combining (11) and (12), our final approximation of (10) is

$$\pi_i \mid \boldsymbol{\pi}_{(i)}, \mu, \tau, \theta, \boldsymbol{s}, \phi = 0 \sim \text{GenBeta}(a_i, b_i, c_i, d_i). \quad (13)$$

It follows from (13) that

$$E_r(\pi_i \mid \boldsymbol{\pi}_{(i)}, \mu, \tau, \theta, \boldsymbol{s}, \phi = 0) \approx c_i + (d_i - c_i) E_u(\pi_i \mid \mu, \tau, \boldsymbol{s})$$

and

$$\begin{aligned} &\text{Var}_r(\pi_i \mid \boldsymbol{\pi}_{(i)}, \mu, \tau, \theta, \boldsymbol{s}, \phi = 0) \\ &\qquad \approx (d_i - c_i)^2 \text{Var}_u(\pi_i \mid \mu, \tau, \boldsymbol{s}), \qquad (14) \end{aligned}$$

where u refers to the unrestricted model and r restricted model. Note that when $c_i = 0$ and $d_i = 1,$ we get $E_r(\pi_i \mid \cdot) = E_u(\pi_i \mid \cdot)$ and $\text{Var}_r(\pi_\ell \mid \cdot) = \text{Var}_u(\pi_\ell \mid \cdot).$ Generally though the estimates of π_i will be a bit different from one scenario to the other. It is also interesting that $\text{Var}_r(\pi_i \mid \cdot) \leq \text{Var}_u(\pi_i \mid \cdot)$ at least approximately. Thus, the restriction $\sum_{i=1}^{\ell} \omega_i \pi_i = \theta$ will reduce variability, when the π_i are estimated. This is true because the $\pi_i, i =$ $1, ..., \ell,$ belong to an $\ell - 1$ dimensional simplex in the ℓ dimensional hypercube while for the unrestricted model $\pi_i, i = 1, ..., \ell,$ belong to the ℓ dimensional hypercube. We expect the largest gain in precision when θ is completely specified, followed by the case when μ_0 is specified and $\tau_0 >> 2,$ and the least gain in precision when $\theta \sim \text{Uniform}(0, 1).$

2.3 Computation

We show how to draw samples from the unrestricted and restricted models. For the unrestricted model we are able to draw random samples from (7) without using Markov chain Monte Carlo methods. However, for the restricted model we use the griddy Gibbs sampler (Ritter and Tanner 1992) to draw samples from (9).

2.3.1 Unrestricted model

We collapse over the $\pi_i,$ draw samples from $p(\mu, \tau \mid \boldsymbol{s})$ using random draws from a bivariate grid, and finally obtain samples from the Rao-Blackwellized densities $\pi_i \mid \mu, \tau, \boldsymbol{s}.$

Then,

$$\pi_i \mid \mu, \tau, \boldsymbol{s} \overset{\text{ind}}{\sim} \text{Beta}\{s_i + \mu\tau,\ f_i + (1-\mu)\tau\},\ i = 1, \ldots, \ell, \quad (15)$$

and integrating out $\boldsymbol{\pi}$, we get

$$p(\mu, \tau \mid \boldsymbol{s}) \propto \prod_{i=1}^{\ell} \frac{B\{s_i + \mu\tau,\ f_i + (1-\mu)\tau\}}{B\{\mu\tau,\ (1-\mu)\tau\}} \times \frac{1}{(1+\tau)^2},$$

$0 < \mu < 1, \tau > 0.$ Letting $\delta = \tau / \tau + 1,$ we have

$$p(\mu, \delta \mid \boldsymbol{s}) \propto \left[\prod_{i=1}^{\ell} \frac{B\{s_i + \mu\tau,\ f_i + (1-\mu)\tau\}}{B\{\mu\tau,\ (1-\mu)\tau\}}\right]_{\tau = \frac{\delta}{1-\delta}},\ 0 < \mu, \delta < 1.$$

First we draw $\mu, \delta \mid \boldsymbol{s}$ using a bivariate grid on $(0, 1)^2$ to obtain a sample of $M \approx 10{,}000$ values of $(\mu^{(h)}, \delta^{(h)})$, $h = 1, \ldots, M$, $\tau^{(h)} = \delta^{(h)}/1 - \delta^{(h)}$. Then we perform a data augmentation in (15) to obtain $\pi^{(h)}$, $h = 1, 2, \ldots, M$, using a composition method. That is, we simply draw $\pi_i \sim \text{Beta}\{s_i + \mu^{(h)}\tau^{(h)}),\ f_i + (1-\mu^{(h)})\tau^{(h)}\},\ i = 1, \ldots, \ell,\ h = 1, \ldots, M.$

To perform the bivariate grid method for sampling from the posterior density of (μ, δ), we divide the interval $(0, 1)$ into 100 sub-intervals; so there are 10,000 little squares in the original unit square. We obtain the heights of the posterior density (without the normalization constant) at the center of each of the 10,000 squares. Because these little squares have the same area, the heights of the bivariate density are proportional to the posterior probabilities that (μ, δ) fall in each of these squares. Thus, we have constructed a joint posterior probability mass function of (μ, δ) on very fine grids. It is easy to draw a sample from the discrete bivariate probability mass function by using the cumulative distribution method. This is actually a random draw of one of the 10,000 squares with probabilities proportional to the heights of the little squares. Then within the selected square we choose a point at random by drawing two uniform random variables (*i.e.*, uniform random jittering). Indeed, this is a very accurate random draw from the joint posterior density of (μ, δ). We draw $M = 10{,}000$ samples from this approximation for posterior inference in a standard Monte Carlo procedure with independent samples, not a Markov chain. Because of the random jittering the numbers are different with probability one.

2.3.2 Restricted model

We show how to draw samples from the restricted model using the Gibbs sampler. The joint conditional posterior density of $\pi_1, \ldots, \pi_{\ell-1}$ is

$$p(\pi_1, \ldots, \pi_{\ell-1} \mid \mu, \tau, \theta, \boldsymbol{s}, \phi = 0) \propto \prod_{i=1}^{\ell-1} \left\{\pi_i^{s_i + \mu\tau - 1}(1-\pi_i)^{f_i + (1-\mu)\tau - 1}\right\} \times \left(\theta - \sum_{i=1}^{\ell-1}\omega_i\pi_i\right)^{s_\ell + \mu\tau - 1} \left\{\sum_{i=1}^{\ell-1}\omega_i\pi_i - \theta + \omega_\ell\right\}^{f_\ell + (1-\mu)\tau - 1} \quad (16)$$

where

$$\theta < \omega_\ell,\ \theta - \omega_\ell < \sum_{i=1}^{\ell-1}\omega_i\,\pi_i < \theta,\ \pi_\ell = \frac{\theta - \sum_{i=1}^{\ell-1}\omega_i\,\pi_i}{\omega_\ell}.$$

Thus, we would obtain samples of $\pi_1, \ldots, \pi_{\ell-1}$ and we set

$$\pi_\ell = \frac{\left(\theta - \sum_{i=1}^{\ell-1}\omega_i\,\pi_i\right)}{\omega_\ell}$$

to complete the vector $\pi_1, \ldots, \pi_\ell$. That is, the constraint is obtained exactly. The conditional posterior density of θ is

$$p(\theta \mid \boldsymbol{\pi}_{(\ell)}, \mu, \tau, \boldsymbol{s}, \phi = 0) \propto \left\{\theta - \sum_{i=1}^{\ell-1}\omega_i\,\pi_i\right\}^{s_\ell + \mu\tau - 1} \left\{\omega_\ell + \sum_{i=1}^{\ell-1}\omega_i\,\pi_i - \theta\right\}^{f_\ell + (1-\mu)\tau - 1} \times \theta^{\mu_0\tau_0 - 1}(1-\theta)^{(1-\mu_0)\tau_0 - 1}, \quad (17)$$

where

$$\sum_{i=1}^{\ell-1}\omega_i\,\pi_i < \theta < \omega_\ell + \sum_{i=1}^{\ell-1}\omega_i\,\pi_i.$$

The joint conditional posterior density of μ and τ is

$$p(\mu, \tau \mid \boldsymbol{\pi}_{(\ell)}, \theta, \boldsymbol{s}, \phi = 0) \propto \frac{q^{\mu\tau}\, r^{(1-\mu)\tau}}{[B(\mu\tau,\ (1-\mu)\tau)]^{\ell}} \times \frac{1}{(1+\tau)^2}, \quad (18)$$

$0 < \mu < 1,\ \tau > 0,\ q = \prod_{i=1}^{\ell}\pi_i,\ r = \prod_{i=1}^{\ell}(1-\pi_i).$

To perform the Gibbs sampler, we need to draw samples from (16), (17) and (18), each in turn, until convergence. We draw μ, τ from $p(\mu, \tau \mid \boldsymbol{\pi}_{(\ell)}, \theta, \boldsymbol{s})$ in a manner similar to drawing from $p(\mu, \tau \mid \boldsymbol{\pi}_{(\ell)})$ in the unrestricted model. It is more difficult to draw sample from (16) and (17). However, we use essentially the same method to draw samples from the conditional posterior density of π_i, $i = 1, \ldots, \ell - 1$, obtained from (16) and θ from (17) which are both proportional to the product of two density functions, one is a truncated beta density and the other a generalized beta density. We next develop some theory to draw a sample from such a density. For this purpose, we state and prove Lemma 2 and Theorem 2.

The density function of interest is

$$f(x) = Af_1(x)\,f_2(x),\ 0 \le c < x < d \le 1, \qquad (19)$$

where

$$f_1(x) = \frac{x^{g-1}(1-x)^{h-1}}{\int_c^d x^{g-1}(1-x)^{h-1}dx},\ c < x < d,\ g, h > 0, \qquad (20)$$

$$f_2(x) = (x-c)^{a-1}(d-x)^{b-1} / \{(d-c)^{a+b-1}B(a,b)\},$$

$$c < x < d,\ a, b > 1, \qquad (21)$$

and, of course,

$$A = 1 \Big/ \int_c^d f_1(x)\,f_2(x)\,dx. \qquad (22)$$

It is worth noting that we are not assuming $g, h > 1$. If this was the case, then $f_1(x)$ and $f_2(x)$ will be both logconcave, thereby making $f(x)$ logconcave, and in this case one can draw a sample from $f(x)$ using the adaptive rejection sampler (ARS, Gilks and Wild 1992). We are providing a specialized algorithm to draw a sample from $f(x)$ which is not logconcave. Even if $f_1(x)$ was logconcave (*i.e.*, $g, h > 1$) this specialized algorithm will still be better than the ARS because the ARS is a general purpose algorithm; see Robert and Casella (1999, page 59). Our algorithm requires less computation and does not need logconcavity; even if there is logconcavity the ARS can perform poorly in the tails of the density function.

Lemma 2 *Consider the density functions* $f_1(x)$ *and* $f_2(x)$ *with* $a, b > 1$.

(a) *Then*

$$\sup_{c<x<d} f_2(x) = \frac{\delta^{a-1}(1-\delta)^{b-1}}{(d-c)\,B(a,b)}, \delta = (a-1) / (a+b-2).$$

(b) *For any* $g > 0, h > 0$ *there exist two constants* H_1 *and* H_2 *such that*

$$0 < H_1 \le A^{-1} \le H_2 < \infty.$$

A proof of Lemma 2 is given in Appendix A.

Theorem 2 *Let* $F_{g,h}(\cdot)$ *be the cdf of* Beta(g, h) *random variable and* $F_{g,h}^{-1}(\cdot)$ *be its inverse. Let*

$$U, V \overset{\text{ind}}{\sim} \text{Uniform}(0, 1),$$

and let

$$X = F_{g,h}^{-1}\{UF_{g,h}(d) + (1-U)\,F_{g,h}(c)\}.$$

If for two real numbers $a, b > 1$,

$$V \le \frac{1}{(d-c)^{a+b-2}}\left(\frac{X-c}{\delta}\right)^{a-1}\left(\frac{d-X}{1-\delta}\right)^{b-1},$$

where $\delta = (a-1) / (a+b-2)$, *then* X *has the density* $f(x) = Af_1(x)\,f_2(x)$.

A proof of Theorem 2 is given in Appendix A.

Theorem 1 gives us the following algorithm for drawing samples from $f(\pi) \propto \pi^{g-1}(1-\pi)^{h-1}(\pi-c)^{a-1}(d-\pi)^{b-1}$, $c < \pi < d,\ g, h > 0,\ a, b > 1$.

Algorithm

(a) Draw $U \sim \text{Uniform}(0, 1)$ and set

$$\pi = F_{g,h}^{-1}\{UF_{g,h}(d) + (1-U)\,F_{g,h}(c)\}.$$

(b) Draw $V \sim \text{Uniform}(0, 1)$. If

$$V \le \frac{1}{(d-c)^{a+b-2}}\left(\frac{\pi-c}{\delta}\right)^{a-1}\left(\frac{d-\pi}{1-\delta}\right)^{b-1},$$

accept π, otherwise go to (a).

Because the binomial sample sizes are arranged in increasing order, in any application it will be true that $a, b > 1$ and $g, h > 0$ (possibly greater than 1 as well). Thus, the algorithm will work. Indeed, in all our examples (one presented here) and simulation exercises the algorithm runs very quickly.

Now, we show how to draw $\pi_i, i = 1, ..., \ell$, and θ. For π_i,

$$p(\pi_i \mid \boldsymbol{\pi}_{(i,\ell)}, \theta, \mu, \tau, \boldsymbol{s}, \phi = 0)$$

$$\propto \pi_i^{a_i-1}(1-\pi_i)^{b_i-1}(\pi_i - c_i)^{b_\ell-1}(d_i - \pi_i)^{a_\ell-1}, c_i < \pi_i < d_i,$$

where $\boldsymbol{\pi}_{(i,\ell)}$ is the vector containing the elements of $\boldsymbol{\pi}$ except for π_i and π_ℓ, and $a_i = s_i + \mu\tau$, $b_i = f_i + (1-\mu)\,\tau$, $i = 1, ..., \ell$,

$$c_i = \left(\theta - \sum_{j=1,\ j\ne i}^{\ell-1} \omega_i\,\pi_i - \omega_\ell\right) \Big/ \omega_i,$$

$$d_i = \left(\theta - \sum_{j=1,\ j\ne i}^{\ell-1} \omega_i\,\pi_i\right) \Big/ \omega_i,\ i = 1, ..., \ell - 1.$$

Apply the theorem to $p(\pi_i \mid \boldsymbol{\pi}_{(i)}, \theta, \mu, \tau, \boldsymbol{s})$, $a_\ell > 1$, $b_\ell > 1$, $i = 1, ..., \ell - 1$.

For θ, we have

$$p(\theta \mid \pi, \mu, \tau, \boldsymbol{s}, \phi = 0)$$

$$\propto \theta^{\mu_0\tau_0-1}(1-\theta)^{(1-\mu_0)\tau_0-1}(\theta-\tilde{c})^{a_\ell-1}(\tilde{d}-\theta)^{b_\ell-1}, \tilde{c} < \theta < \tilde{d},$$

where

$$\tilde{c} = \sum_{i=1}^{\ell-1}\omega_i\,\pi_i, \; \tilde{d} = \omega_\ell + \sum_{i=1}^{\ell-1}\omega_i\,\pi_i.$$

Again, apply the theorem, $a_\ell > 1$, $b_\ell > 1$.

When θ is fully specified (*i.e.*, θ is not random), we do not have to draw θ. However, when $\theta \sim \text{Uniform}(0, 1)$ a priori $(\mu_0 = 1/2, \tau_0 = 2)$, we have a simplification. In this case,

$$\theta \mid \pi_{(\ell)}, \mu, \tau, \boldsymbol{s}, \phi = 0 \sim \text{GenBeta}(a_\ell, b_\ell, \tilde{c}, \tilde{d})$$

and $\theta = \tilde{c} + (\tilde{d} - \tilde{c})X$, where $X \sim \text{Beta}(a_\ell, b_\ell)$, has the required density.

For both the unrestricted and restricted models we use 10,000 iterates to make posterior inference about the binomial probabilities, π_i. Under the unrestricted model these are simply random draws and no monitoring is required. For the restricted model, running the griddy Gibbs sampler, we drew 11,000 iterates, used 1,000 as a "burn in" (a conservative number because convergence occur much earlier as evident in the trace plots) and we found negligible correlations among the iterates. Thus, we used 10,000 iterates to make inference about the binomial probabilities. For both the unrestricted and the three restricted models it takes only a few seconds on our 2×833 MHz alpha computer.

3. Numerical studies

In Section 3.1 we describe an illustrative example to show the main features of the restriction. In Section 3.2 we describe a simulation study to show frequentist properties of the Bayes estimators, and we show deeper insight into the differences among the four scenarios. Note again that when we performed the computations, it is convenient to order the domain sizes so that the largest domain comes last.

3.1 Illustrative example

We have used data in the third National Health and Nutrition Examination (NHANES) Survey to illustrate our method. We have studied body mass index for teenagers, and we have data on the sample obtained. The domains (small areas) are formed by crossing ethnicity (white, black, Mexican) and sex (male, female). We have separated out the teenagers with respect to whether they were in middle school or high school at the time of the survey. Thus, there are 12 small domains. The data are presented in the first four columns of Table 1 by domain. Note that domains MWM, MBF, MWF and HBF are relatively sparse with 4, 2, 5, 5 obese teenagers respectively; for the twelve domains the sample consists of 959 with 130 obese teenagers (*i.e.*, the overall proportion of obese individuals is 0.136 approximately). In column 4 of Table 1 we have also presented the direct estimates by domains, and these estimates range from 0.069 to 0.228. The estimates for the smallest domains will be unreliable. Moreover, when the beta-binomial models are used, these estimates will regress to the overall sample mean of 0.136, creating a possible bias. Our method is expected to increase precision beyond the unrestricted model because the restricted model uses more information about the weighted sum. Clearly, predictors based on either the restricted model or the beta-binomial model are biased if the specified model is wrong.

We have taken $\mu_0 = 0.136$, the overall sample proportion, and $\tau_0 = 959$, the total sample size. Less optimistic choices can be used. For example, $\tau_0 = 100$, say; but this choice makes very little difference. However, it is worth noting that using the observed data to specify the prior distribution can artificially decrease the posterior variance. Typically a survey practitioner will have an appropriate specification from a prior survey or a census. One cannot specify values for μ_0 and τ_0 which are completely out of line and will create huge biases. Here τ_0 is a prior sample size and μ_0 is a prior mean of θ. This method permits a sensible value for θ; we are essentially adding a degree of uncertainty about knowledge of the linear combination. Thus, these specifications are not unreasonable.

We have applied our method as described for the four scenarios. In the other columns of Table 1 we study the estimates of the small area probabilities. We present the posterior mean (PM), posterior standard deviation (PSD),

$$\text{RMSE} = \sqrt{(\hat{\pi} - \text{PM})^2 + \text{PSD}^2},$$

where $\hat{\pi}$ is the direct estimate, and the 95% highest posteriori density (HPD) interval (Int). As is expected, the PSDs are roughly in the increasing order: Model 2, Model 3, Model 4 and Model 1; in some cases the differences are important. The PMs for Models 1, 2 and 3 are mostly similar, but for Model 4 the PMs are mostly smaller than the other three models. There is much improvement of Models 2 and 3 over Model 1 at least in terms of precision. This gain becomes less important for Model 4, the model with the greatest uncertainty about θ.

Table 1
Comparison of the four models using posterior mean (PM), posterior standard deviation, root mean square error (RMSE), and 95% credible HPD intervals (Int) of π_i by domain (D) for the NHANES data

D	s	n	$\hat{\pi}$	PM	PSD	RMSE	Int	PM	PSD	RMSE	Int
				Model 1				Model 2			
1	4	47	0.085	0.114	0.033	0.044	(0.051, 0.179)	0.111	0.032	0.041	(0.049, 0.170)
2	2	29	0.069	0.112	0.037	0.057	(0.042, 0.183)	0.111	0.036	0.055	(0.041, 0.178)
3	10	44	0.227	0.175	0.044	0.068	(0.100, 0.264)	0.177	0.041	0.065	(0.108, 0.260)
4	5	62	0.081	0.107	0.030	0.040	(0.047, 0.159)	0.107	0.027	0.038	(0.054, 0.160)
5	10	74	0.135	0.134	0.030	0.030	(0.077, 0.194)	0.134	0.028	0.028	(0.080, 0.190)
6	12	69	0.174	0.158	0.036	0.039	(0.089, 0.227)	0.155	0.031	0.036	(0.095, 0.214)
7	8	79	0.101	0.116	0.028	0.031	(0.065, 0.173)	0.115	0.027	0.030	(0.065, 0.166)
8	5	62	0.081	0.107	0.030	0.040	(0.052, 0.169)	0.105	0.029	0.038	(0.042, 0.153)
9	28	123	0.228	0.196	0.036	0.048	(0.129, 0.262)	0.196	0.032	0.045	(0.131, 0.253)
10	10	111	0.090	0.106	0.026	0.030	(0.059, 0.155)	0.105	0.024	0.028	(0.061, 0.150)
11	16	122	0.131	0.132	0.026	0.026	(0.083, 0.183)	0.130	0.023	0.023	(0.090, 0.179)
12	20	137	0.146	0.144	0.026	0.026	(0.094, 0.194)	0.141	0.022	0.023	(0.100, 0.184)
				Model 3				Model 4			
1	4	47	0.085	0.111	0.033	0.042	(0.044, 0.169)	0.109	0.032	0.040	(0.050, 0.172)
2	2	29	0.069	0.111	0.037	0.056	(0.039, 0.179)	0.108	0.036	0.053	(0.037, 0.173)
3	10	44	0.227	0.175	0.043	0.068	(0.093, 0.260)	0.170	0.044	0.072	(0.091, 0.255)
4	5	62	0.081	0.106	0.029	0.038	(0.050, 0.160)	0.103	0.030	0.038	(0.048, 0.164)
5	10	74	0.135	0.134	0.029	0.029	(0.077, 0.189)	0.129	0.030	0.030	(0.067, 0.184)
6	12	79	0.174	0.156	0.034	0.038	(0.090, 0.217)	0.151	0.036	0.043	(0.087, 0.222)
7	8	69	0.101	0.118	0.028	0.033	(0.062, 0.171)	0.111	0.028	0.029	(0.061, 0.167)
8	5	62	0.081	0.107	0.030	0.040	(0.051, 0.165)	0.102	0.030	0.036	(0.050, 0.159)
9	28	123	0.228	0.195	0.034	0.047	(0.138, 0.265)	0.189	0.035	0.052	(0.123, 0.255)
10	10	111	0.090	0.107	0.024	0.029	(0.062, 0.156)	0.104	0.025	0.029	(0.051, 0.149)
11	16	122	0.131	0.132	0.024	0.024	(0.086, 0.179)	0.126	0.025	0.025	(0.083, 0.179)
12	20	137	0.146	0.143	0.024	0.024	(0.095, 0.191)	0.137	0.025	0.027	(0.091, 0.189)

Note: The four models are: Model 1 - no restriction; Model 2 - fixed θ; Model 3 - informative prior for θ; Model 4 - uniform prior for θ. Domains are formed by crossing school (middle school - M, high school - H), race (white - W, black - B, mexican american - M) and sex (male - M, female - F). Thus, the domains are: 1-MWM, 2-MBF, 3-MMM, 4-MWF, 5-MBM, 6-MMF, 7-HWM, 8-HBF, 9-HMM, 10-HWF, 11- HBM, 12-HMF (*e.g.*, the first domain consists of middle school white boys). n is the number of teenagers and s the number of obese teenagers in each domain. Data are taken from the 35 largest counties in the US. An estimate of the overall probability is $130 / 959 \approx 0.136,$ and for the first domain $\hat{p} = 4 / 47 = 0.085;$ the numerical standard errors are all smaller than 0.001; $\text{RMSE} = \sqrt{(\hat{\pi} - \text{PM})^2 + \text{PSD}^2}.$

We also study very briefly the nuisance parameter θ. We note that the weighted average of the direct estimators of the small areas is 0.136 (more accurately 0.1355599). When θ is held fixed at 0.1355599, the weighted average of the posterior means is 0.136. When θ has the informative prior, the weighted average of the posterior means is 0.136, and for θ the PM is 0.136, the PSD is 0.008, and a 95% HPD interval for θ is $(0.122, 0.152)$. When θ has the uniform prior, the weighted average of the posterior means is 0.132, and for θ the PM is 0.131, the PSD is 0.011, and a 95% HPD interval for θ is $(0.110, 0.151)$. This shows the deficiencies of the uniform prior which we use only for comparison. It is worth noting that $\mu_1, \ldots, \mu_{\ell-1}$ and θ are computed first. Then μ_ℓ is obtained by subtraction. This is done at each iterate of the Gibbs sampler. Then, the posterior summaries for $\sum_{i=1}^{\ell} \omega_i \pi_i$ and θ are computed. So there will be very minor discrepancies which are due to rounding.

Finally, we have selected the four smallest domains to compare the posterior densities of the probabilities. We have used the Parzen-Rosenblatt kernel density estimator to estimate the posterior densities; see Silverman (1986) for details. Figure 1 compares the estimated posterior densities for the four models. It is interesting that as the domain sizes increase, the four models get closer together. Also, for all cases the tails of the distributions in each panel are very similar; the differences in these distributions though lie in the modal intervals (*i.e.*, interval containing the mode), and their heights. As expected, the posterior density corresponding to the unrestricted model is the shortest, simply because it has more variability. Model 4 has posterior density shifted to the left and is slightly bimodal for the smallest domain. Thus, inference about the modes of these distributions will be different. But inference involving the tails will not be so different; except for Model 4, 95% credible intervals will be similar.

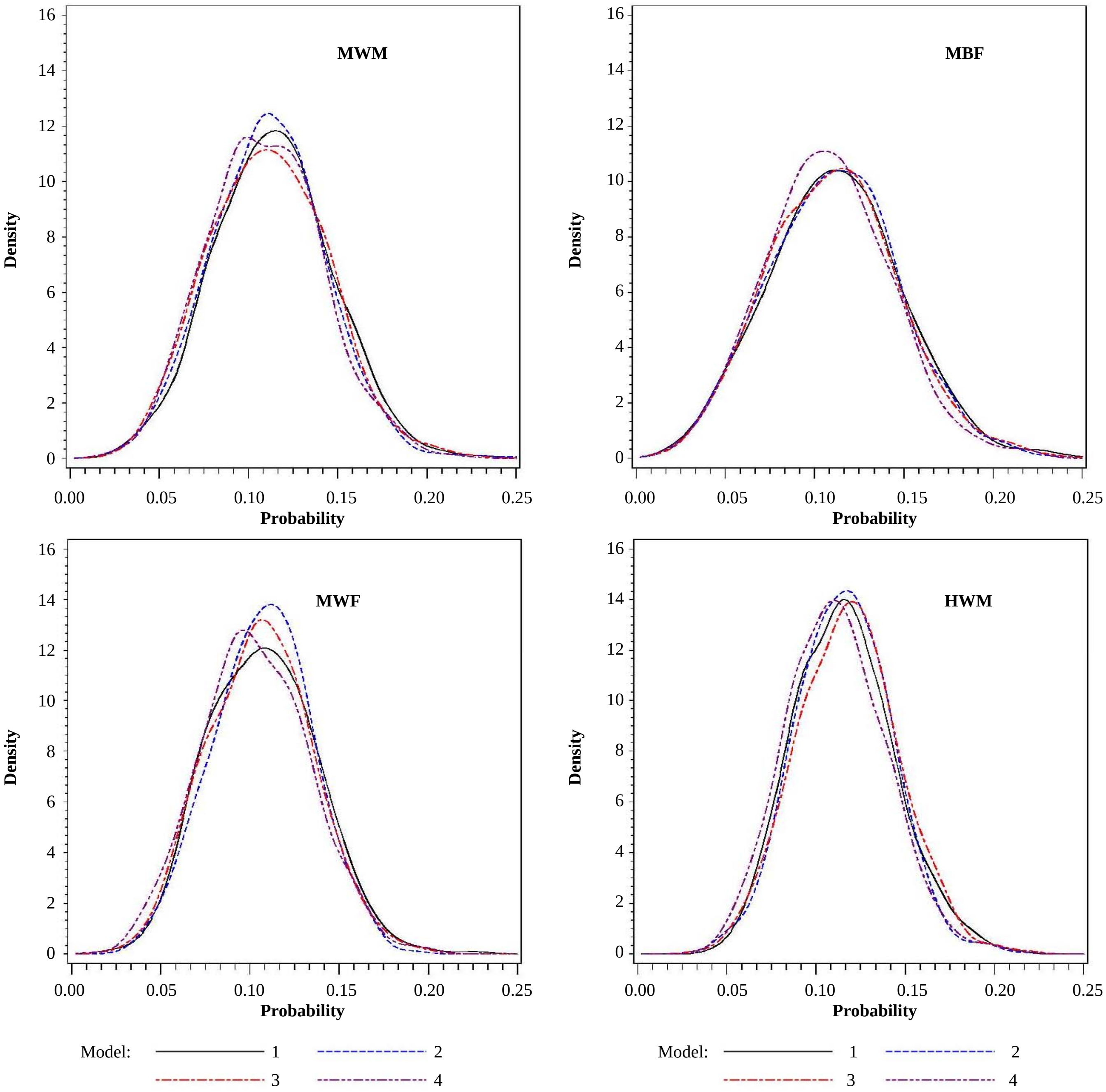


Figure 1 Plots of the estimated posterior densities of π_1, π_2, π_4, and π_7 for the four models and NHANES data

3.2 Simulation study

We use a simulation study to assess the statistical properties of our method. We want to see if the gain in precision persists and to see how the estimators of the probabilities are shifted. We also study the frequentist properties of the estimators of the probabilities. In the description of the simulation it is convenient to use the abbreviated names of the models which are UR (Model 1, no restriction), FI (Model 2, fixed θ), IN (Model 3, informative prior for θ) and UN (Model 4, uniform prior for θ).

We set $\theta_0 = 0.15$, $\mu_0 = \theta_0$ and $\tau_0 = 100$. We have selected three values of $\ell = 12, 24, 36$, 12 being the number of areas in the NHANES data. We drew the sample sizes from a uniform density in (25, 150), again to reflect the NHANES data. First, we generated

$$\pi_i \overset{\text{iid}}{\sim} \text{Beta}\{\mu_0\tau_0, (1-\mu_0)\tau_0\}, \; i = 1, ..., \ell.$$

To do this latter task, we drew sets of $\ell\ \pi_i$ until $\theta_0 - w_\ell \le \sum_{i=1}^{\ell-1}\omega_i\,\pi_i \le \theta_0$; set $\pi_\ell = (\theta_0 - \sum_{i=1}^{\ell-1}\omega_i\,\pi_i)\ /\ w_\ell$. Then, we generated

$$s_i \overset{\text{ind}}{\sim} \text{Binomial}(n_i,\ \pi_i).$$

We have generated 1,000 data sets in this manner for each of $\ell = 12, 24, 36$. Then, we fit the four models (one unrestricted and three restricted models). The process is very fast (*i.e.*, for samples sizes of 12, 24, 30 there were respectively 22, 90, 153 rejects in the 1,000 samples). We fit each data set using random samples for the unrestricted model and the griddy Gibbs sampler for the restricted models. We fit the 1,000 data sets in a couple of hours on our on our 2×833 MHz alpha computer.

For these 1,000 simulations we study PM, the coverage (C), the bias (B), PSD, RMSE and width (W) of the 95% credible intervals. For each domain we compute the bias $\text{PM} - \pi$, then we average these values over all domains and simulation runs, and this quantity we now call B. Associated with B we also computed AB, the average of $|\ \text{PM} - \pi\ |$. Similarly, we have computed

$$\text{RMSE} = \sqrt{(\text{PM} - \pi)^2 + \text{PSD}^2}$$

for each domain and each simulation run and we average these over all domains and simulation runs. Note that the true probabilities, π_i, are known by design. We obtain the coverage (C) by computing the proportion of all intervals containing the true value of π_i over all domains and simulation runs. We also obtain the average of the widths of the 95% credible intervals. Numerical standard errors are obtained for all quantities.

In Table 2 we study the estimates of the small area probabilities. It is convenient to use the shorter names of the four models for our discussion. For IN the PMs are close to the nominal value of 0.15, but for UN the PMs are smaller than the nominal value particularly for UN at $\ell = 12$. We observe that the coverage for all the models UR, FI and UN are always larger than the nominal value of 95%, but for model IN these coverages are smaller than the nominal value of 95%. A similar difference exists for the bias; while the bias is small for all models, models UR, FI (the specified value of θ is 0.15) and UN have negative biases but IN has positive bias. Except for $\ell = 36$ IN has the largest AB. The PSDs are mostly similar and the RMSEs share the same features; there are some differences at $\ell = 12$. The four models get similar as ℓ increases; when ℓ is large there appears to be no need for our method. However, again the gain in precision appears to be in the increasing order FI, IN, UN and UR.

In most applications the exact value of θ is unknown. Therefore, the PSDs of the π_i, under the situation where θ is assumed known, are likely to underestimate the true PSDs. So we study the deviations of the PSDs of IN and UN from those of FI, and we compute the ratios, $R_1 = \text{PSD}_{\text{IN}}\ /\ \text{PSD}_{\text{FI}}$ and $R_2 = \text{PSD}_{\text{UN}}\ /\ \text{PSD}_{\text{FI}}$. In Table 3 we present the five-number summaries of these ratios by sample size. Most of the ratios are around 1 (*i.e.*, inter-quartile range) with some tendency for them to be larger than 1. (Note that the maxima at $\ell =$ and $\ell = 24$ are outliers possibly due to bad simulated samples.) Thus, overall the PSDs under IN and UN are not much larger under FI.

Table 2
Simulation: Comparison of the four models using coverage (C), bias and average absolute bias (B and AB), posterior standard deviation (PSD), root mean squared error (RMSE) and width of the 95% credible intervals (W) of π_i

ℓ	Model	C	B	AB	PSD	RMSE	W
12	UR	$0.960_{0.0018}$	$-0.002_{0.0003}$	$0.0231_{0.00016}$	$0.033_{0.0001}$	$0.043_{0.0001}$	$0.125_{0.0003}$
	FI	$0.961_{0.0018}$	$-0.000_{0.0003}$	$0.0219_{0.00020}$	$0.031_{0.0001}$	$0.040_{0.0001}$	$0.118_{0.0003}$
	IN	$0.946_{0.0021}$	$0.005_{0.0003}$	$0.0275_{0.00066}$	$0.032_{0.0001}$	$0.043_{0.0001}$	$0.122_{0.0002}$
	UN	$0.956_{0.0019}$	$-0.000_{0.0003}$	$0.0261_{0.00019}$	$0.032_{0.0001}$	$0.042_{0.0001}$	$0.122_{0.0003}$
24	UR	$0.957_{0.0013}$	$-0.001_{0.0002}$	$0.0229_{0.00012}$	$0.031_{0.0000}$	$0.041_{0.0001}$	$0.119_{0.0002}$
	FI	$0.957_{0.0013}$	$-0.000_{0.0002}$	$0.0224_{0.00013}$	$0.030_{0.0000}$	$0.040_{0.0001}$	$0.116_{0.0002}$
	IN	$0.943_{0.0015}$	$0.006_{0.0002}$	$0.0252_{0.00058}$	$0.030_{0.0000}$	$0.041_{0.0001}$	$0.116_{0.0001}$
	UN	$0.952_{0.0014}$	$-0.000_{0.0002}$	$0.0236_{0.00012}$	$0.031_{0.0002}$	$0.041_{0.0002}$	$0.118_{0.0005}$
36	UR	$0.960_{0.0010}$	$-0.001_{0.0001}$	$0.0224_{0.00009}$	$0.030_{0.0000}$	$0.040_{0.0001}$	$0.117_{0.0001}$
	FI	$0.961_{0.0010}$	$-0.000_{0.0001}$	$0.0218_{0.00009}$	$0.030_{0.0000}$	$0.039_{0.0001}$	$0.115_{0.0001}$
	IN	$0.948_{0.0012}$	$0.005_{0.0002}$	$0.0224_{0.00009}$	$0.030_{0.0000}$	$0.040_{0.0001}$	$0.114_{0.0001}$
	UN	$0.957_{0.0011}$	$-0.000_{0.0001}$	$0.0228_{0.00010}$	$0.030_{0.0000}$	$0.040_{0.0001}$	$0.116_{0.0001}$

Note: The four models are: Model 1 - no restriction (UR); Model 2 - fixed θ (FI); Model 3 - informative prior for θ (IN); Model 4 - uniform prior for θ (UN). $\text{RMSE} = \sqrt{(\pi - \text{PM})^2 + \text{PSD}^2}$. The notation a_b means a is an estimate and b is the standard error.

Table 3
Simulation: A study of the posterior standard deviation (PSD) of the π_i using five number summaries of the ratios, R_1 and R_2, by sample size

ℓ	Ratio	Min	Q_1	Med	Q_3	Max
12	R_1	0.673	0.972	1.032	1.091	5.329
	R_2	0.022	0.984	1.034	1.086	85.370
24	R_1	0.019	0.965	1.005	1.047	16.017
	R_2	0.024	0.979	1.014	1.049	486.960
36	R_1	0.690	0.962	0.998	1.034	1.236
	R_2	0.837	0.979	1.011	1.044	1.243

Note: $R_1 = \text{PSD}_{\text{IN}} / \text{PSD}_{\text{FI}}$ and $R_2 = \text{PSD}_{\text{UN}} / \text{PSD}_{\text{FI}}$. The five summaries are minimum (min), first quartile (Q_1), median (med), third quartile (Q_3) and maximum (max).

In Table 4 we study the estimate of θ for the two pertinent models IN and UN. For both models the coverage probabilities are smaller than the nominal value, and the coverage for UN is smaller than the interval for IN. Bias is small for both models, positive for IN and negative for UN. Except for $\ell = 36$ IN has by far the larger AB. The PSDs and RMSEs are generally smaller for IN, and the widths of the 95% credible intervals are significantly smaller for IN. It appears that it is difficult to estimate θ under UN, but IN appears to be somewhat better.

In Table 5 we present more detailed result (*i.e.*, by domain) for the case when the number of domains is 12. To show further gains in precision, we have reduced the sample size to half as much [*i.e.*, we drew the sample sizes uniformly in the interval (12, 75)]. We present the posterior standard deviation and the posterior root mean square error, averaged over the simulation runs. Again the standard errors are presented. We note that all the probability contents (not presented) are at least the nominal value of 95%. The numerical standard errors are small in all cases. The PSDs and RMSEs are in the right order. Note that because the sample sizes are arranged in order from smallest to largest, there is a decrease in the PSDs and RMSEs as the domain numbers go up.

Table 4
Simulation: Comparison of the informative (IN) and the uniform (UN) models using posterior mean (PM), coverage (C), bias and average absolute bias (B and AB), posterior standard deviation (PSD), root mean squared error (RMSE) and width of the 95% credible intervals (W) of π_i

ℓ	Model	PM	C	B	AB	PSD	RMSE	W
12	IN	$0.149_{0.0012}$	$0.853_{0.0112}$	$0.000_{0.0003}$	$0.00152_{0.00081}$	$0.008_{0.0000}$	$0.012_{0.0002}$	$0.030_{0.0001}$
	UN	$0.138_{0.0005}$	$0.881_{0.0102}$	$-0.012_{0.0004}$	$0.00038_{0.00003}$	$0.011_{0.0001}$	$0.016_{0.0002}$	$0.042_{0.0002}$
24	IN	$0.153_{0.0015}$	$0.833_{0.0118}$	$0.003_{0.0015}$	$0.00212_{0.00103}$	$0.007_{0.0006}$	$0.012_{0.0015}$	$0.024_{0.0015}$
	UN	$0.145_{0.0029}$	$0.842_{0.0115}$	$-0.005_{0.0003}$	$0.00012_{0.00006}$	$0.008_{0.0001}$	$0.012_{0.0002}$	$0.030_{0.0002}$
36	IN	$0.150_{0.0002}$	$0.828_{0.0119}$	$0.000_{0.0002}$	$0.00004_{0.00000}$	$0.004_{0.0000}$	$0.007_{0.0001}$	$0.017_{0.0001}$
	UN	$0.145_{0.0003}$	$0.794_{0.0128}$	$-0.005_{0.0002}$	$0.00009_{0.00000}$	$0.006_{0.0000}$	$0.010_{0.0001}$	$0.024_{0.0001}$

Note: The two models considered are: Model 3 – informative prior for θ and model 4 - uniform prior for θ. RMSE = $\sqrt{(\theta_0 - \text{PM})^2 + \text{PSD}^2}$. The notation a_b means a is an estimate and b is the standard error.

Table 5
Simulation: Comparison of the four models using posterior standard deviation and root mean square error (RMSE) of π_i by domain (D)

	Unrestricted		Fixed		Informative		Uniform	
D	PSD	RMSE	PSD	RMSE	PSD	RMSE	PSD	RMSE
1	$0.048_{0.0003}$	$0.057_{0.0004}$	$0.046_{0.0003}$	$0.054_{0.0004}$	$0.045_{0.0002}$	$0.056_{0.0005}$	$0.047_{0.0004}$	$0.056_{0.0005}$
2	$0.046_{0.0003}$	$0.055_{0.0004}$	$0.044_{0.0003}$	$0.053_{0.0004}$	$0.044_{0.0002}$	$0.054_{0.0005}$	$0.045_{0.0004}$	$0.054_{0.0005}$
3	$0.044_{0.0002}$	$0.053_{0.0004}$	$0.042_{0.0002}$	$0.050_{0.0004}$	$0.042_{0.0002}$	$0.052_{0.0005}$	$0.043_{0.0003}$	$0.051_{0.0004}$
4	$0.042_{0.0002}$	$0.050_{0.0004}$	$0.040_{0.0002}$	$0.047_{0.0004}$	$0.040_{0.0002}$	$0.050_{0.0004}$	$0.041_{0.0002}$	$0.049_{0.0004}$
5	$0.041_{0.0002}$	$0.049_{0.0004}$	$0.038_{0.0002}$	$0.046_{0.0004}$	$0.039_{0.0002}$	$0.048_{0.0004}$	$0.039_{0.0003}$	$0.048_{0.0005}$
6	$0.040_{0.0002}$	$0.048_{0.0004}$	$0.037_{0.0002}$	$0.045_{0.0004}$	$0.037_{0.0002}$	$0.048_{0.0004}$	$0.038_{0.0003}$	$0.047_{0.0005}$
7	$0.038_{0.0002}$	$0.046_{0.0004}$	$0.035_{0.0002}$	$0.043_{0.0003}$	$0.036_{0.0002}$	$0.046_{0.0004}$	$0.037_{0.0003}$	$0.045_{0.0004}$
8	$0.037_{0.0002}$	$0.045_{0.0003}$	$0.034_{0.0002}$	$0.041_{0.0003}$	$0.036_{0.0002}$	$0.046_{0.0004}$	$0.036_{0.0003}$	$0.044_{0.0004}$
9	$0.036_{0.0002}$	$0.044_{0.0003}$	$0.033_{0.0002}$	$0.040_{0.0004}$	$0.034_{0.0001}$	$0.044_{0.0004}$	$0.035_{0.0003}$	$0.042_{0.0004}$
10	$0.035_{0.0002}$	$0.043_{0.0003}$	$0.032_{0.0002}$	$0.039_{0.0003}$	$0.034_{0.0001}$	$0.044_{0.0004}$	$0.034_{0.0003}$	$0.042_{0.0004}$
11	$0.034_{0.0001}$	$0.042_{0.0003}$	$0.031_{0.0002}$	$0.038_{0.0003}$	$0.033_{0.0001}$	$0.042_{0.0004}$	$0.033_{0.0003}$	$0.041_{0.0004}$
12	$0.035_{0.0002}$	$0.047_{0.0005}$	$0.031_{0.0002}$	$0.042_{0.0004}$	$0.034_{0.0003}$	$0.047_{0.0006}$	$0.034_{0.0007}$	$0.046_{0.0008}$

Note: The four models are: Model 1 - no restriction; Model 2 - fixed θ; Model 3 - informative prior for θ; Model 4 - uniform prior for θ. RMSE= $\sqrt{(\pi_i - \text{PM})^2 + \text{PSD}^2}$. The notation a_b means a is an estimate and b is the standard error. Here 12 domains are used and the original simulated sample sizes are divided by 2.

We study the posterior density of π_1 for $\ell = 12$, and we compare the four models. Again we use the Parzen-Rosenblatt density estimator. In Figure 2 we present the estimated posterior densities (Parzen-Rosenblatt) averaged over the 1,000 runs for $\ell = 12$. We obtain the same results as for the BMI data. Again the tails are similar. FI is the tallest density and UN is the shortest. UN is slightly shifted to the left of IN. In Figure 3 we present a systematic sample of 10 densities from the 1,000 simulation runs by model. We can see large variation among the 10 estimated posterior densities. Again we can see that FI is tallest; UR, FI and UN show similar variation with IN slightly taller. Thus, it is important to take the average for comparison as in Figure 2.

4. Concluding remarks

We have extended the beta-binomial model of small area estimation to accommodate a prior specification of a weighted average of the area probabilities. We have used the Bayesian approach which is particularly attractive for problems with awkward likelihood functions as in our application with the constraint of the weighted average of the beta-binomial model. We viewed the constraint as prior knowledge which can be precise or less informative. The griddy Gibbs sampler is used to fit the models, thereby avoiding the more sophisticated Metropolis-Hastings sampler. We have developed a theory which permits sampling from a density function which is proportional to the product of a truncated beta-binomial density and a generalized beta density. We have found that overall our complete algorithm forming the griddy Gibbs sampler runs efficiently and fast.

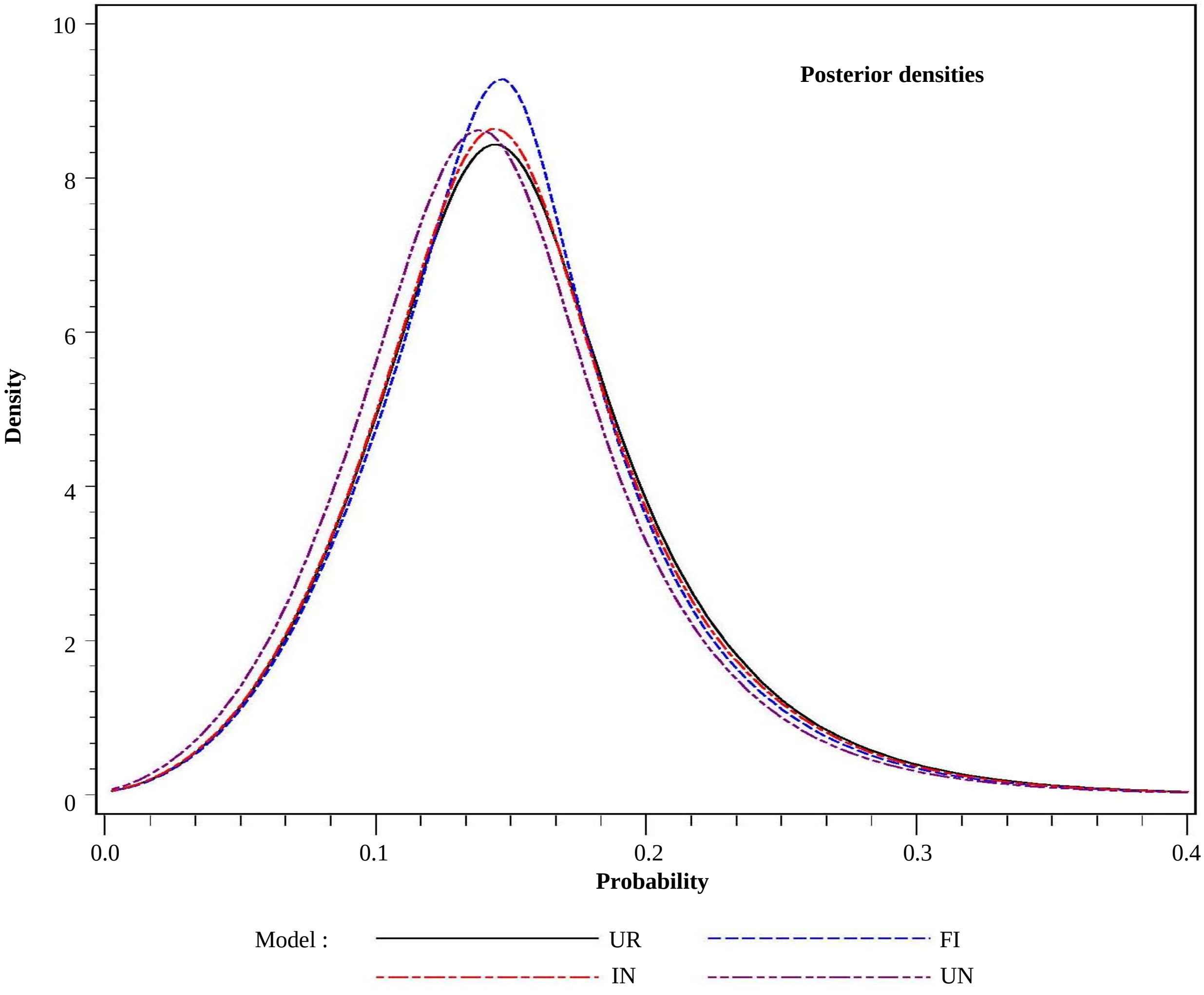


Figure 2 Plots of the estimated posterior densities of π_1 by model when there are 12 domains

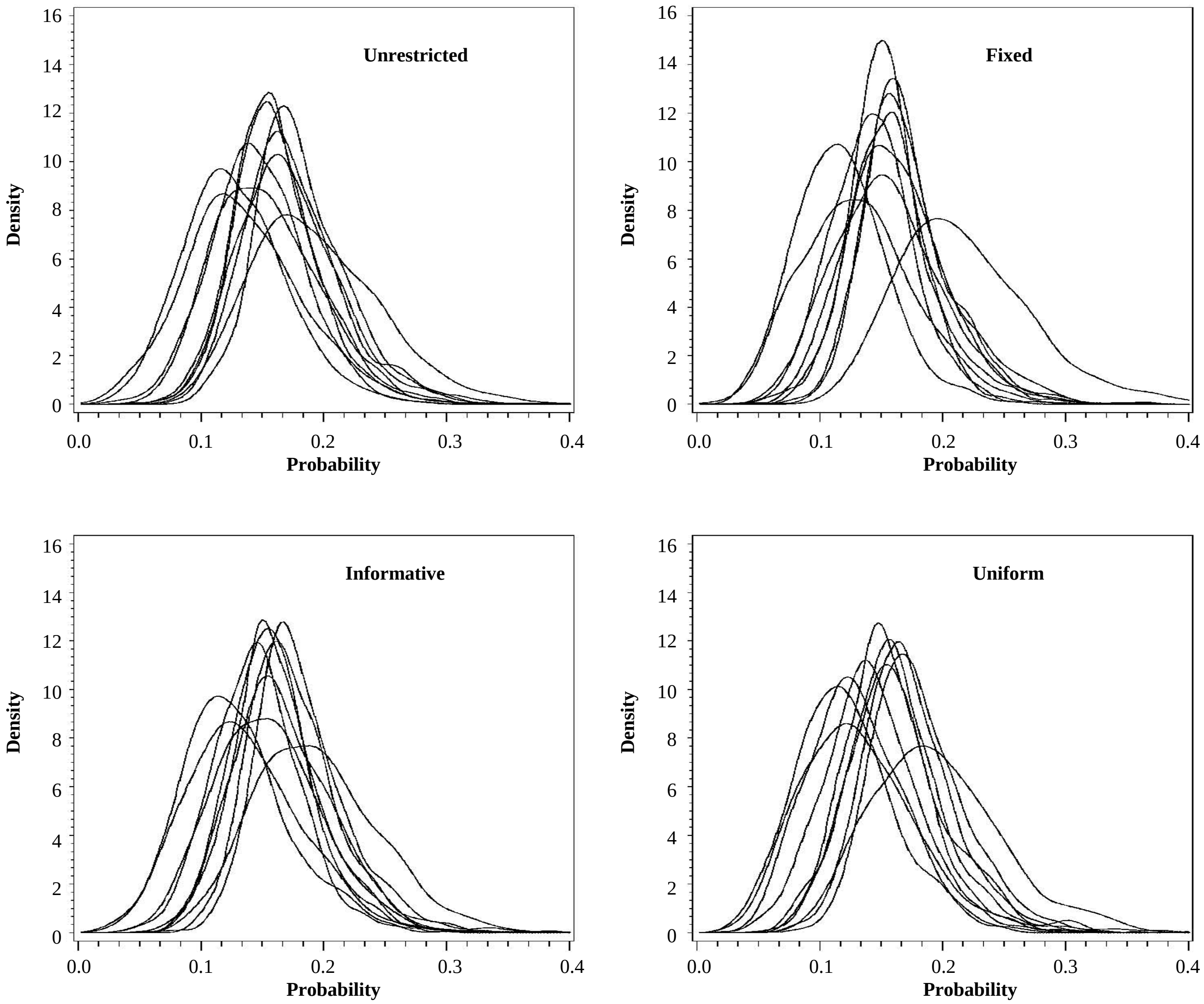


Figure 3 Plots of the estimated posterior densities of π_1 for a systematic sample of size 10 from the 1,000 runs by model when there are 12 domains

We have shown that there could be gains in precision when extra information is incorporated into the beta-binomial model. We have considered three scenarios in which a survey practitioner (a) can not specify any constraint (standard beta-binomial model for small areas), (b) can specify a constraint and the parameter completely, and (c) can specify a constraint and information which can be used to construct a prior distribution for the parameter. Our example on obesity of children in the National Health and Nutrition Examination Survey and simulation study showed that the gain in precision beyond (a) is in an order with (b) larger than (c). As the exact algebraic arguments are difficult, we obtained an analytical approximation which shows that indeed there could be gain in precision of (b) over (a). For comparison we have considered a fourth scenario in which θ has vague information, and as expected, it turned to be rather uninteresting and inefficient.

It is straight forward to make Bayesian predictive inference about the finite population mean of each small area. Let $P_i = T_i / N_i$ denote the finite population proportion for the i^{th} area, where $T_i = \sum_{j=1}^{N_i} y_{ij}$, y_{ij} are the binary responses, and N_i, the number of individuals in the i^{th} area, is assumed known. Now $T_i = t_i^{(s)} + t_i^{(ns)}$, where $t_i^{(s)}$ and $t_i^{(ns)}$ are respectively the sample total and the nonsample total. Now under any of the models $t_i^{(ns)} \mid \pi_i \sim \text{Binomial}(n_i, \pi_i)$ and $p(t_i^{(ns)} \mid \boldsymbol{y}_s) = \int p(t_i^{(ns)} \mid \pi_i)\, p(\pi \mid \boldsymbol{y}_s)\, d\pi_i$, where $\boldsymbol{y}_s = (y_1, \ldots, y_\ell)'$. Thus, it is easy to obtain the empirical posterior density of P_i using a sampling-based method. Nandram and Sedransk (1993) obtained some analytical features of P_i when τ is known, but not with the constraint; see also Nandram (1998).

We mention a generalization of our restricted beta-binomial hierarchical Bayesian model to the Dirichlet-multinomial model (*e.g.*, Nandram 1998). Let $\boldsymbol{y}_i$ be c-vector of cell counts (*i.e.*, number of people possessing one of c traits), and let n_i denote the sample sizes within the i^{th} area, $i = 1, \ldots, \ell$. We assume

$$\boldsymbol{y}_i \mid \pi_i \overset{\text{ind}}{\sim} \text{Multinomial}(n_i, \boldsymbol{\pi}_i), \boldsymbol{\pi}_i \mid \boldsymbol{\mu}, \tau, \boldsymbol{\theta} \overset{\text{iid}}{\sim} \text{Dirichlet}(\boldsymbol{\mu}\tau)$$

with $\sum_{i=1}^{\ell} w_i \boldsymbol{\pi}_i = \boldsymbol{\theta}$. Finally $\boldsymbol{\theta} \sim \text{Dirichlet}(\boldsymbol{\mu}_0 \tau_0)$, where $\boldsymbol{\mu}_0$ and τ_0 are to be specified, and independently $p(\boldsymbol{\mu}, \tau) = (k-1)!/(1+\tau)^2, 0 < \mu_k < 1, k = 1, \ldots, c, \sum_{k=1}^{c} \mu_k = 1$. With k constraints this problem is much more complex, but we plan to work on it. Other extensions to nonignorable nonresponse (Nandram and Choi 2002) and two-way categorical tables are possible.

Acknowledgements

The authors are grateful to the Associate Editor and the two referees who helped enormously to improve the quality of the presentation.

Appendix A

Proofs of lemmas 1, 2 and theorems 1, 2

Proof of lemma 1

This is a special case of a general result. Using the multiplication rule and because the prior is proper, it is clear that the joint density of $\boldsymbol{\pi}, \mu, \tau, \boldsymbol{s}$ "integrates" to one. Therefore, the joint posterior density of $\boldsymbol{\pi}, \mu, \tau$ given $\boldsymbol{s}$ is proper.

Proof of theorem 1

Let $\mathcal{T} = \{(\pi, \mu, \tau, \theta)\text{: } 0 < \pi_i < 1,\ i = 1, \ldots, \ell,\ 0 < \mu < 1, \tau > 0, \theta - \omega_\ell \le \sum_{i=1}^{\ell-1} \omega_i \pi_i \le \theta,\ 0 < \theta < 1,\ \pi_\ell = (\theta - \sum_{i=1}^{\ell-1} \omega_i \pi_i) / \omega_\ell\}$ and $\mathcal{T}^* = \{(\pi, \mu, \tau)\text{: } 0 < \pi_i < 1,\ i = 1, \ldots, \ell, 0 < \mu < 1,\ \tau > 0\}$; note that $\mathcal{T} \subset \mathcal{T}^*$.

Let $\tilde{g}(\pi, \mu, \tau \mid \boldsymbol{s})$ denote the right-hand side of the unrestricted posterior density in (7) and $\tilde{p}(\boldsymbol{\pi}_{(\ell)}, \mu, \tau, \theta \mid \boldsymbol{s}, \phi = 0)$ denote the right-hand side of the restricted posterior density in (9). Noting that $\pi_\ell = (\theta - \sum_{i=1}^{\ell-1} \omega_i \pi_i) / \omega_\ell$, we observe that

$$\tilde{p}(\boldsymbol{\pi}_{(\ell)}, \mu, \tau, \theta \mid \boldsymbol{s}, \phi = 0) =$$

$$\tilde{g}(\pi, \mu, \tau \mid \boldsymbol{s}) \times \theta^{\mu_0 \tau_0 - 1} (1 - \theta)^{(1-\mu_0)\tau_0 - 1},\ (\pi, \mu, \tau, \theta) \in \mathcal{T}.$$

Because $\theta^{\mu_0 \tau_0 - 1} (1 - \theta)^{(1-\mu_0)\tau_0 - 1}$ is proportional to the density function of beta random variable, we have

$$\int_{\mathcal{T}} \tilde{p}(\boldsymbol{\pi}_{(\ell)}, \mu, \tau, \theta \mid s, \phi = 0)\, d\boldsymbol{\pi} d\mu d\tau d\theta =$$

$$A \int_{\mathcal{T}} \tilde{g}(\pi, \mu, \tau \mid s)\, d\boldsymbol{\pi} d\mu d\tau \le A \int_{\mathcal{T}^*} \tilde{g}(\pi, \mu, \tau \mid \boldsymbol{s}) d\boldsymbol{\pi} d\mu d\tau,$$

where $A = B\{\mu_0 \tau_0, (1 - \mu_0)\tau_0\}$ is the beta function. By lemma 1, $\int_{\mathcal{T}^*} \tilde{g}(\pi, \mu, \tau \mid \boldsymbol{s})\, d\boldsymbol{\pi} d\mu d\tau < \infty$. Thus, $p(\boldsymbol{\pi}_{(\ell)}, \mu, \tau, \theta \mid \boldsymbol{s}, \phi = 0)$ is proper.

Proof of Lemma 2 (*a*)

This can be proved in two ways. The second derivative of $\log\{f_2(x)\}$ is negative in (c, d), and so the first derivative, when set to zero, provides a unique mode which is $\delta d + (1 - \delta)c$. Alternatively, because $(X - c)/(d - c) \sim \text{Beta}(a, b)$ with $a, b > 1$, there is a unique mode for $(X - c)/(d - c)$, and this translates to $\delta d + (1 - \delta)c$; note that $\delta d + (1 - \delta)c$ is a point in (c, d). Thus, substituting $\delta d + (1 - \delta)c$ into $f_2(x)$, we have

$$\sup_{c<x<d} f_2(x) = \delta^{a-1}(1 - \delta)^{b-1} / (d - c)\, B(a, b).$$

Proof of Lemma 2 (*b*)

Because $a, b > 1$, $x \ge x - c$ and $1 - x \ge d - x$, it is true that

$$A^{-1} \ge D^{-1} \int_c^d (x - c)^{a+g-2} (d - x)^{b+h-2} dx,$$

where $D = (d - c)^{a+b-1} B(a, b) B(g, h)\{F_{g,h}(d) - F_{g,h}(c)\}$ and $F_{g,h}(x)$ is the cdf of a standard beta random variable in $(0, 1)$. Note that because $c < d$ (strictly) and $F_{g,h}(x)$ is monotone increasing in $(0, 1)$, $F_{g,h}(d) - F_{g,h}(c) > 0$ (strictly). By comparison with the generalized beta density [*i.e.*, $\text{Beta}(a + g - 1, b + h - 1, c, d)$], the integral is $(d - c)^{a+b+g+h-3} B(a + g - 1, b + h - 1)$. Thus,

$$A^{-1} \ge \frac{(d - c)^{g+h-2} B(a + g - 1, b + h - 1)}{B(a, b) B(g, h)\{F_{g,h}(d) - F_{g,h}(c)\}} = H_1 > 0.$$

Also, we have

$$A^{-1} \le \int_c^d f_1(x) \sup_{c<x<d} f_2(x)\, dx.$$

Then by Lemma 2 (a),

$$A^{-1} \le \frac{\delta^{a-1}(1 - \delta)^{b-1}}{(d - c) B(a, b)} \int_c^d f_1(x)\, dx = \frac{\delta^{a-1}(1 - \delta)^{b-1}}{(d - c) B(a, b)}$$

$$= H_2 < \infty.$$

Proof of theorem 2

To show the claim, we calculate the cdf $F_X(\cdot)$ of the random variable X defined in the Theorem. We have

$$\begin{aligned} F_X(x) &= P(X \le x) \\ &= P[F_{g,h}^{-1}\{U F_{g,h}(d) + (1 - U) F_{g,h}(c)\} \le x] \\ &= P[U F_{g,h}(d) + (1 - U) F_{g,h}(c) \le F_{g,h}(x)] \\ &= P[U\{F_{g,h}(d) - F_{g,h}(c)\} \le F_{g,h}(x) - F_{g,h}(c)] \\ &= P\left[U \le \frac{F_{g,h}(x) - F_{g,h}(c)}{F_{g,h}(d) - F_{g,h}(c)}\right]. \end{aligned}$$

Now, since $U \sim \text{Uniform}(0, 1)$, from the above expression for $F_X(\cdot)$, we have $F_X(x) = 1$ if $x \geq d$ and $F_X(x) = 0$ if $x \leq c$. When $c \leq x \leq d$, we have

$$F_X(x) = \frac{F_{g,h}(x) - F_{g,h}(c)}{F_{g,h}(d) - F_{g,h}(c)}.$$

This shows that X has the truncated beta density $f_1(x)$ in (20).

Now, looking to use the accept-reject algorithm, consider

$$\frac{f(x)}{f_1(x)} = Af_2(x).$$

By Lemma 2, we have

$$\sup_{c<\pi<d} \left\{ \frac{f(x)}{f_1(x)} \right\} = A \sup_{c<\pi<d} f_2(x) = A \frac{\delta^{a-1}(1-\delta)^{b-1}}{(d-c)B(a, b)} < \infty.$$

Thus, by the accept-reject algorithm, if

$$V \leq \frac{1}{(d-c)^{a+b-2}} \left(\frac{X-c}{\delta} \right)^{a-1} \left(\frac{d-X}{1-\delta} \right)^{b-1},$$

then X has the density $f(x)$ in (19).

References

Cochran, W.G. (1977). *Sampling Techniques*, third edition. New York: John Wiley & Sons, Inc.

Gilks, W.R., and Wild, P. (1992). Adaptive rejection sampling for gibbs sampling. *Journal of the Royal Statistical Society*, Series C, 41, 337-348.

Gelman, A. (2006). Prior distribution for variance parameters in hierarchical models. *Bayesian Analysis*, 1, 515-533.

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.

Hillmer, S.C., and Trabelsi, A. (1987). Benchmarking of economic time series. *Journal of the American Statistical Association*, 82, 1064-1071.

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

Nandram, B. (1998). A Bayesian analysis of the three-stage hierarchical multinomial model. *Journal of Statistical Computation and Simulation*, 61, 97-126.

Nandram, B., and Choi, J.W. (2002). Hierarchical Bayesian nonresponse models for binary data from small areas with uncertainty about ignorability. *Journal of the American Statistical Association*, 97, 381-388.

Nandram, B., and Choi, J.W. (2002). A Bayesian analysis of a proportion under non-ignorable non-response. *Statistics in Medicine*, 21, 9, 1189-1212.

Nandram, B., and Sedransk, J. (1993). Bayesian predictive inference for a finite population proportion: Two-stage cluster sampling. *Journal of the Royal Statistical Society*, Series B, 55, 399-408.

Nandram, B., Toto, M.C.S. and Choi, J.W. (2011). A Bayesian benchmarking of the Scott-Smith model for small areas. *Journal of Statistical Computation and Simulation* (in press, preprint).

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

Ritter, C., and Tanner, M.A. (1992). The gibbs sampler and the griddy gibbs sampler. *Journal of the American Statistical Association*, 87, 861-868.

Robert, C.P., and Casella, G. (1999). *Monte Carlo Statistical Methods*. New York: Springer-Verlag.

Silvapulle, M.J., and Sen, P.K. (2006). *Constrained Statistical Inference*: *Inequality*, *Order and Shape Restrictions*. New York: John Wiley & Sons, Inc.

Silverman, B.W. (1986). *Density Estimation*. London: Chapman and Hall.

Survey Methodology, December 2011
Vol. 37, No. 2, pp. 153-170
Statistics Canada, Catalogue No. 12-001-X

On bias-robust mean squared error estimation for pseudo-linear small area estimators

Ray Chambers, Hukum Chandra and Nikos Tzavidis [1]

Abstract

We propose a method of mean squared error (MSE) estimation for estimators of finite population domain means that can be expressed in pseudo-linear form, *i.e.*, as weighted sums of sample values. In particular, it can be used for estimating the MSE of the empirical best linear unbiased predictor, the model-based direct estimator and the M-quantile predictor. The proposed method represents an extension of the ideas in Royall and Cumberland (1978) and leads to MSE estimators that are simpler to implement, and potentially more bias-robust, than those suggested in the small area literature. However, it should be noted that the MSE estimators defined using this method can also exhibit large variability when the area-specific sample sizes are very small. We illustrate the performance of the method through extensive model-based and design-based simulation, with the latter based on two realistic survey data sets containing small area information.

Key Words: Best linear unbiased prediction; M-quantile model; Model-based direct estimation; Random effects model; Small area estimation.

1. Introduction

Linear models, and linear predictors based on these models, are widely used in survey-based inference. However, such models run the risk of misspecification, particularly with regard to second order and higher moments. Bias-robust methods for estimating the mean squared error (MSE) of linear predictors of finite population quantities, *i.e.*, methods that remain approximately unbiased under failure of assumptions about second order and higher moments, have been developed. Valliant, Dorfman and Royall (2000, Chapter 5) discuss bias-robust MSE estimation for such predictors when a population is assumed to follow a linear model.

In this paper we address a subsidiary problem, which is that of bias-robust MSE estimation for estimators of finite population domain means that can be expressed in pseudo-linear form, *i.e.*, as weighted sums, but where the weights can depend on the sample values of the variable of interest. An important application, and one that motivates our approach, is small area inference. Consequently from now on we use ‘area’ to refer to a domain of interest. Our approach represents an extension of the ideas in Royall and Cumberland (1978) and appears to lead to simpler to implement MSE estimators than those that have been suggested in the small area literature.

The structure of the paper is as follows. In section 2 we discuss MSE estimation under an area-specific linear model. That is, we focus on estimation of the conditional MSE. We then show how our approach can be used for estimating the MSE of three different small area linear predictors when they are expressed in pseudo-linear form, (a) the empirical best linear unbiased predictor or EBLUP (Henderson 1953); (b) the model-based direct estimator (MBDE) of Chandra and Chambers (2009); and (c) the M-quantile predictor (Chambers and Tzavidis 2006). In section 3 we present results from a series of simulation studies that illustrate the model-based and the design-based properties of our approach to MSE estimation. Finally, in section 4 we summarize our main findings. Throughout, we use either i or h to index the D small areas of interest, and either j or k to index the distinct population units in these areas.

2. Bias-robust MSE estimation for pseudo-linear estimators

2.1 MSE estimation under an area-specific linear model

We consider the situation where we have a finite population of size N from which a sample of size n is drawn. We assume that this population consists of D non-overlapping domains, each one of which contains sampled units, with small realised sample sizes in each of the sampled domains. As noted earlier, and following standard practice, we refer to these domains as areas from now on. We assume also that there is a known number N_i of population units in area i, with n_i of these sampled. The total number of units in the population is $N = \sum_{i=1}^{D} N_i$, with corresponding total sample size $n = \sum_{i=1}^{D} n_i$. In what follows, we use s to denote the collection of units in sample, with s_i the subset drawn from area i, and use expressions like $j \in i$ and $j \in s$ to refer to the units making up area i and sample s respectively.

[1] Ray Chambers, Centre for Statistical and Survey Methodology, University of Wollongong, Wollongong, NSW, 2522, Australia. E-mail: ray@uow.edu.au; Hukum Chandra, Indian Agricultural Statistics Research Institute, Library Avenue, New Delhi-110012, India. E-mail: hchandra@iasri.res.in; Nikos Tzavidis, Social Statistics and Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, SO17 1BJ, UK. E-mail: n.tzavidis@soton.ac.uk.

Linear models are often used to motivate estimators for population means. However, when estimates are required for the corresponding area means, it is usually not realistic to assume that a linear model that applies to the population as a whole also applies within each area. We therefore adopt a conditional approach, and consider MSE estimation for estimators of area means when different linear models apply within different areas. In particular, we focus on estimators that can be expressed as weighted sums of the sample values, referring to them as 'linear' in what follows to indicate that they have a linear structure.

To start, let y_j denote the value of *Y* for unit *j* of the population and suppose that this unit is in area *i*. We also assume an area-specific linear model for y_j of the form

$$y_j = \boldsymbol{x}_j^T \beta_i + e_j. \qquad (1)$$

Here $\boldsymbol{x}_j$ is a $p \times 1$ vector of unit level auxiliary variables for unit *j*, β_i is a $p \times 1$ vector of area-specific regression coefficients and e_j is a unit level random effect with mean zero and variance σ_j^2 that is uncorrelated between different population units. We do not make any assumptions about σ_j^2 at this point. Note that throughout this paper we assume that the sampling method used is non-informative for the population values of *Y* given the corresponding values of the auxiliary variables and knowledge of the area affiliations of the population units. As a consequence, (1) applies at both sample and population level.

Let $\boldsymbol{y}_s$ denote the column vector of sample values of y_j and let $\boldsymbol{w}_{is} = \{w_{ij};\ j \in s\}$ denote the column vector of fixed weights such that $\hat{m}_i = \boldsymbol{w}_{is}^T \boldsymbol{y}_s = \sum_{j \in s} w_{ij} y_j$ is a linear estimator of $m_i = N_i^{-1} \sum_{j \in i} y_j$. By 'fixed' here we mean that these weights do not depend on the sample values of *Y*. Moreover, we assume $w_{ij} = O(n_i^{-1})$ for $j \in s_i, w_{ij} = o(n_i^{-1})$ for $j \notin s_i$, and $\sum_{j \in s} w_{ij} = 1$. Here s_i denotes the n_i sample units from area *i*. The bias of $\hat{m}_i$ under (1) is then

$$E(\hat{m}_i - m_i) = \left(\sum_{h=1}^{D} \sum_{j \in s_h} w_{ij} \boldsymbol{x}_j^T \beta_h\right) - \bar{\boldsymbol{x}}_i^T \beta_i, \qquad (2)$$

where $\bar{\boldsymbol{x}}_i$ denotes the vector of average values of the auxiliary variables in area *i*. Similarly, the prediction variance of $\hat{m}_i$ under (1) is

$$\mathrm{Var}(\hat{m}_i - m_i) = N_i^{-2} \left\{\sum_{h=1}^{D} \sum_{j \in s_h} a_{ij}^2 \sigma_j^2 + \sum_{j \in r_i} \sigma_j^2\right\}, \quad (3)$$

where r_i denotes the non-sampled units in area *i* and $a_{ij} = N_i w_{ij} - I(j \in i)$. We use $I(A)$ to denote the indicator function for event *A*, so $I(j \in i)$ takes the value 1 if population unit *j* is from area *i* and is zero otherwise. Note that since a_{ij} is $O(N_i n_i^{-1})$ for $j \in s_i$, the first term within the braces in (3) is the leading term of this prediction variance if N_i is large compared to n_i.

Let $j \in h$. We consider the important special case where $\mu_j = E(y_j \mid \boldsymbol{x}_j) = \boldsymbol{x}_j^T \beta_h$ is estimated by $\hat{\mu}_j = \boldsymbol{x}_j^T \hat{\beta}_h = \sum_{k \in s} \phi_{kj} y_k$, with the ϕ_{kj} corresponding to suitable weights. Then

$$y_j - \hat{\mu}_j = (1 - \phi_{jj}) y_j - \sum_{k \in s(-j)} \phi_{kj} y_k$$

and so

$$\mathrm{Var}(y_j - \hat{\mu}_j) = \sigma_j^2 \left\{(1 - \phi_{jj})^2 + \sum_{k \in s(-j)} \phi_{kj}^2 (\sigma_k^2 / \sigma_j^2)\right\} \quad (4)$$

under (1). Here $s(-j)$ denotes the sample s with unit j excluded. If in addition $\hat{\mu}_j$ is unbiased for μ_j under (1), *i.e.*,

$$E(y_j - \hat{\mu}_j) = 0, \qquad (5)$$

we can then adopt the approach of Royall and Cumberland (1978) and estimate (3) by

$$\hat{V}(\hat{m}_i) = N_i^{-2} \left\{\sum_{h=1}^{D} \sum_{j \in s_h} a_{ij}^2 \hat{\lambda}_j^{-1} (y_j - \hat{\mu}_j)^2 + \sum_{j \in r_i} \hat{\sigma}_j^2\right\}, \quad (6)$$

where $\hat{\lambda}_j = (1 - \phi_{jj})^2 + \sum_{k \in s(-j)} \hat{\gamma}_{kj} \phi_{kj}^2$ and $\hat{\gamma}_{kj} = \hat{\sigma}_k^2 / \hat{\sigma}_j^2$. Usually, the estimates $\hat{\sigma}_j^2$ of the residual variances in (6) are derived under a 'working model' refinement to (1). In the situation of most concern to us, where the sample sizes within the different areas are too small to reliably estimate area-specific variability, a pooling assumption can be made, *i.e.*, $\sigma_j^2 = \sigma^2$, in which case we put

$$\hat{\sigma}_j^2 = \hat{\sigma}^2 = n^{-1} \sum_{j \in s} \left\{(1 - \phi_{jj})^2 + \sum_{k \in s(-j)} \phi_{kj}^2\right\}^{-1} (y_j - \hat{\mu}_j)^2.$$

In this case (6) becomes

$$\hat{V}(\hat{m}_i) = N_i^{-2} \sum_{j \in s} \left\{a_{ij}^2 + (N_i - n_i) n^{-1}\right\} \hat{\lambda}_j^{-1} (y_j - \hat{\mu}_j)^2, \quad (7)$$

where now $\hat{\lambda}_j = (1 - \phi_{jj})^2 + \sum_{k \in s(-j)} \phi_{kj}^2$. Since any assumptions regarding σ_j^2 in the working model extension of (1) only affect second order terms in (3), the estimator (7) is bias-robust, *i.e.*, it remains approximately unbiased under misspecification of the second order moments of this working model.

A corresponding estimator of the MSE of $\hat{m}_i$ under (1) follows directly. This is

$$\hat{M}(\hat{m}_i) = \hat{V}(\hat{m}_i) + \hat{B}^2(\hat{m}_i), \qquad (8)$$

where

$$\hat{B}(\hat{m}_i) = \sum_{h=1}^{D} \sum_{j \in s_h} w_{ij} \hat{\mu}_j - N_i^{-1} \sum_{j \in i} \hat{\mu}_j \qquad (9)$$

is the obvious unbiased estimator of (2).

Use of the square of the unbiased estimator (9) of the bias of $\hat{m}_i$ in the conditional MSE estimator (8) can be criticised because this term is not itself unbiased for the squared bias term in MSE. This can be corrected by replacing (9) by

$$\hat{M}(\hat{m}_i) = \hat{V}(\hat{m}_i) + \hat{B}^2(\hat{m}_i) - \hat{V}\{\hat{B}(\hat{m}_i)\}, \tag{10}$$

where $\hat{V}\{\hat{B}(\hat{m}_i)\}$ is a suitable estimator of the variance of (9). However, we do not recommend use of (10). To see this, let $\bar{\boldsymbol{\beta}} = D^{-1}\sum_{h=1}^{D}\hat{\boldsymbol{\beta}}_h$ and put $\boldsymbol{d}_h = \hat{\boldsymbol{\beta}}_h - \bar{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}_h$ is the estimator of $\boldsymbol{\beta}_h$ implied by the weights ϕ_{kj}. Furthermore, put $w_{hi} = \sum_{j \in s_h} w_{ij}$ and $\bar{\boldsymbol{x}}_{whi} = w_{hi}^{-1}\sum_{j \in s_h} w_{ij}\boldsymbol{x}_j$, so $\bar{\boldsymbol{x}}_{wi} = \sum_{h=1}^{D}\sum_{j \in s_h} w_{ij}\boldsymbol{x}_j = \sum_{h=1}^{D} w_{hi}\bar{\boldsymbol{x}}_{whi}$ is the estimate of $\bar{\boldsymbol{x}}_i$ based on the weights w_{ij}. Finally, let $\delta_{hi} = \bar{\boldsymbol{x}}_h^T\boldsymbol{d}_h - \bar{\boldsymbol{x}}_i^T\boldsymbol{d}_i$ and put $\delta_i = \sum_{h=1}^{D} w_{hi}\delta_{hi}$. Then (9) can be written

$$\begin{aligned}
\hat{B}(\hat{m}_i) &= (\bar{\boldsymbol{x}}_{wi} - \bar{\boldsymbol{x}}_i)^T\bar{\boldsymbol{\beta}} \\
&\quad + \left(\sum_{h=1}^{D} w_{hi}\bar{\boldsymbol{x}}_{whi}^T\boldsymbol{d}_h - \bar{\boldsymbol{x}}_i^T\boldsymbol{d}_i\right) \\
&= (\bar{\boldsymbol{x}}_{wi} - \bar{\boldsymbol{x}}_i)^T\bar{\boldsymbol{\beta}} \\
&\quad + \left(\sum_{h=1}^{D} w_{hi}(\bar{\boldsymbol{x}}_{whi} - \bar{\boldsymbol{x}}_h)^T\boldsymbol{d}_h + \sum_{h=1}^{D} w_{hi}\bar{\boldsymbol{x}}_h^T\boldsymbol{d}_h - \bar{\boldsymbol{x}}_i^T\boldsymbol{d}_i\right) \\
&= (\bar{\boldsymbol{x}}_{wi} - \bar{\boldsymbol{x}}_i)^T\bar{\boldsymbol{\beta}} \\
&\quad + \sum_{h=1}^{D} w_{hi}(\bar{\boldsymbol{x}}_{whi} - \bar{\boldsymbol{x}}_h)^T\boldsymbol{d}_h + \delta_i.
\end{aligned} \tag{11}$$

Typically, D will be large and the leading term in the variance of (9) will be the variance of δ_i in (11). If this leading term is large, then $\hat{V}\{\hat{B}(\hat{m}_i)\}$ will also be large, and (10) could take negative values. We therefore recommend that (8), rather than (10), be used. An immediate consequence is that (8) is then a conservative estimator of the MSE of $\hat{m}_i$ under (1). This may be acceptable provided that the variance of δ_i is small. However, for very small values of n_i this variance can be large, causing (8) to substantially overestimate the actual MSE of $\hat{m}_i$. We therefore recommend a preliminary empirical assessment of the size of the variance of δ_i relative to the value of (7) in this situation. If this assessment indicates that the variance of δ_i dominates (7), then (8) should not be used.

2.2 MSE estimation for pseudo-linear small area estimators

The approach to conditional MSE estimation outlined in the previous sub-section assumed that the weights defining the linear estimator $\hat{m}_i$ do not depend on the sample values of *Y*. However, most small area estimators do not satisfy this condition, in the sense that they are pseudo-linear in structure, with weights that do depend on these sample values. For example, the Best Linear Unbiased Predictor (BLUP) of m_i under the linear mixed model variant of (1) where the area-specific regression parameters $\boldsymbol{\beta}_i$ are independent and identically distributed realisations of a random variable with expected value $\boldsymbol{\beta}$ and covariance matrix $\boldsymbol{\Gamma}$, can be written as a weighted sum of the sample values of *Y* where the weights depend on $\boldsymbol{\Gamma}$ (see Royall 1976). Consequently, the empirical version of this predictor, the widely used EBLUP, is computed by substituting an efficient sample estimate of $\boldsymbol{\Gamma}$ (*e.g.*, the REML estimate) into the BLUP weights. If the linear mixed model assumption is true, this sample estimator of $\boldsymbol{\Gamma}$ converges to the true value and consequently the EBLUP weights converge to the BLUP weights. That is, for large values of the overall sample size *n*, we can treat the EBLUP weights as fixed and use the MSE estimator (8) for the EBLUP. Of course, the EBLUP weights are not really fixed, and so (8) is therefore an approximation to the true MSE of the EBLUP that ignores the contribution to this MSE arising from the variability in estimation of $\boldsymbol{\Gamma}$. However, this potential underestimation needs to be balanced against the bias robustness of (8) under misspecification of the second order moments of *Y*.

An important advantage of (8) is that it can be used with a range of small area estimators that can be expressed in pseudo-linear form. In particular, many small area estimators developed under models that are variants of (1) can be written in this form, *i.e.*, as weighted sums of the sample values of *Y*. To illustrate, we now focus on three such estimators: the EBLUP (Rao 2003, Chapter 6), the Model-Based Direct Estimator (MBDE) of Chandra and Chambers (2009) and the M-quantile predictor of Chambers and Tzavidis (2006). Each of these estimators can be written in pseudo-linear form, with weights that satisfy $w_{ij} = O(n_i^{-1})$ for $j \in s_i$ and $w_{ij} = o(n_i^{-1})$ for $j \notin s_i$, and so (8) can be used.

2.2.1 MSE estimation for the EBLUP

We first consider the well-known EBLUP for m_i based on a unit level linear mixed model extension of (1) of the form

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{u}_i + \boldsymbol{e}_i \tag{12}$$

where $\boldsymbol{y}_i$ is the N_i-vector of population values of y_j in area *i*, $\boldsymbol{X}_i$ is the corresponding $N_i \times p$ matrix of auxiliary variable values $\boldsymbol{x}_j$, $\boldsymbol{Z}_i$ is the $N_i \times q$ component of $\boldsymbol{X}_i$ corresponding to the *q* random components of $\boldsymbol{\beta}$, $\boldsymbol{u}_i$ is the associated *q*-vector of area-specific random effects and $\boldsymbol{e}_i$ is the N_i-vector of individual random effects. It is typically assumed that the area and individual effects are mutually independent, with the area effects independently and identically distributed as $N(0, \boldsymbol{\Omega})$ and the individual effects independently and identically distributed as $N(0, \sigma^2)$. See Rao (2003, Chapter 6) for development of the underlying theory of this predictor. We note that the EBLUP can be written in pseudo-linear form,

$$\hat{m}_i^{\text{EBLUP}} = \sum_{j \in s} w_{ij}^{\text{EBLUP}} y_j = (\boldsymbol{w}_{is}^{\text{EBLUP}})^T\boldsymbol{y}_s \tag{13}$$

where

$$\boldsymbol{w}_{is}^{\text{EBLUP}} = (w_{ij}^{\text{EBLUP}})$$

$$= N_i^{-1}[\Delta_{is} + \{\hat{\boldsymbol{H}}_s^T \boldsymbol{X}_r^T + (\boldsymbol{I}_n - \hat{\boldsymbol{H}}_s^T \boldsymbol{X}_s^T)\hat{\Sigma}_{ss}^{-1}\hat{\Sigma}_{sr}\}\Delta_{ir}].$$

Here Δ_{ir} is the vector of size $N - n$ that 'picks out' the non-sampled units in area *i*, $\boldsymbol{X}_s$ and $\boldsymbol{X}_r$ are the matrices of order $n \times p$ and $(N - n) \times p$ respectively of the sample and non-sample values of the auxiliary variables, $\boldsymbol{I}_n$ is the identity matrix of order *n*, $\hat{\boldsymbol{H}}_s = (\boldsymbol{X}_s^T \hat{\Sigma}_{ss}^{-1} \boldsymbol{X}_s)^{-1} \boldsymbol{X}_s^T \hat{\Sigma}_{ss}^{-1}$, $\hat{\Sigma}_{ss} = \hat{\sigma}^2 \boldsymbol{I}_n + \text{diag}\{\boldsymbol{Z}_{is}\hat{\Omega}\boldsymbol{Z}_{is}^T;\, i = 1,\ldots,D\}$ and $\hat{\Sigma}_{sr} = \text{diag}\{\boldsymbol{Z}_{is}\hat{\Omega}\boldsymbol{Z}_{ir}^T;\, i = 1, \ldots, D\}$. Here $\boldsymbol{Z}_{is}(\boldsymbol{Z}_{ir})$ is the sample (non-sample) component of $\boldsymbol{Z}_i$ and $\hat{\sigma}^2$ and $\hat{\Omega}$ are suitable (*e.g.*, ML or REML) estimates of the variance components of (12).

Given this setup, estimation of the conditional MSE of the EBLUP can be carried out using (8) with weights defined following (13). In turn, this requires that we have access to unbiased estimators $\hat{\mu}_j$ of the area specific individual expected values μ_j. However, such estimators may be unstable when area sample sizes are small. Consequently, it is tempting to replace $\hat{\mu}_j$ by the EBLUP for y_j, *i.e.*, $\hat{y}_j^{\text{EBLUP}} = \boldsymbol{x}_j^T \hat{\beta}^{\text{EBLUE}} + \boldsymbol{z}_j^T \hat{\boldsymbol{u}}_i^{\text{EBLUP}}$, where $\hat{\beta}^{\text{EBLUE}}$ denotes the Empirical Best Linear Unbiased Estimator of β in the linear mixed model (12) and $\hat{\boldsymbol{u}}_i^{\text{EBLUP}}$ denotes the predicted area effect for the area *i* that contains observation *j*. Unfortunately, because of the well-known shrinkage effect associated with EBLUPs, this approach is not recommended. To illustrate this, we note that $\hat{V}(\hat{m}_i)$ in (8) uses $(y_j - \hat{\mu}_j)^2$ as an estimator of $E(y_j - \mu_j)^2$. The bias in this estimator is therefore

$$E(y_j - \hat{\mu}_j)^2 - E(y_j - \mu_j)^2$$

$$= -2E(y_j - \mu_j)(\hat{\mu}_j - \mu_j) + E(\hat{\mu}_j - \mu_j)^2$$

$$= -E\{(\hat{\mu}_j - \mu_j)(2y_j - \mu_j - \hat{\mu}_j)\}$$

so we anticipate that $\hat{V}(\hat{m}_i)$ will be negatively biased if $E\{(\hat{\mu}_j - \mu_j)(2y_j - \mu_j - \hat{\mu}_j)\}$ is positive and vice versa. Now let sample unit *j* be from area *i* and consider the special case of a random intercept model for y_j, *i.e.*, $y_j = \boldsymbol{x}_j^T \beta + u_i + e_j$ where u_i is the random effect for area *i* and e_j is a random individual effect uncorrelated with u_i. Here $\mu_j = \boldsymbol{x}_j^T \beta + u_i$. Suppose that we have a large overall sample size, allowing us to replace $\hat{\beta}^{\text{EBLUE}}$ by β. The EBLUP $\hat{\mu}_j = \hat{y}_j^{\text{EBLUP}}$ can then be approximated by $\tilde{\mu}_j = \boldsymbol{x}_j^T \beta + \gamma_i u_i$, where γ_i is a 'shrinkage' factor. It follows that

$$(\tilde{\mu}_j - \mu_j)(2y_j - \mu_j - \tilde{\mu}_j) = 2u_i(\gamma_i - 1)e_i - u_i^2(\gamma_i - 1)^2$$

so $E(y_j - \hat{\mu}_j)^2 - E(y_j - \mu_j)^2 \approx (\gamma_i - 1)^2 \sigma_u^2$. That is, we expect $\hat{V}(\hat{m}_i)$ to be positively biased if we use the shrunken EBLUP $\hat{y}_j^{\text{EBLUP}}$ to define $\hat{\mu}_j$. We also note that this bias disappears (approximately) if we 'unshrink' the residual component of this EBLUP. For example, in the case of the popular random intercepts model, we use

$$\hat{\mu}_j = \boldsymbol{x}_j^T \hat{\beta}^{\text{EBLUE}} + (\bar{y}_{is} - \bar{\boldsymbol{x}}_{is}^T \hat{\beta}^{\text{EBLUE}}) = \bar{y}_{is} + (\boldsymbol{x}_j - \bar{\boldsymbol{x}}_{is})^T \hat{\beta}^{\text{EBLUE}}$$

where $\bar{y}_{is}$ and $\bar{\boldsymbol{x}}_{is}$ denote the sample means of *Y* and *X* respectively in area *i*. Given (12) is the working model, a general expression for such an 'unshrunken' estimator is

$$\hat{\mu}_j = \boldsymbol{x}_j^T \hat{\beta}^{\text{EBLUE}} + \boldsymbol{z}_j^T \tilde{\boldsymbol{u}}_i \tag{14}$$

where $\tilde{\boldsymbol{u}}_i = (\boldsymbol{Z}_{is}^T \boldsymbol{Z}_{is})^{-1} \boldsymbol{Z}_{is}^T (y_{is} - \boldsymbol{X}_{is} \hat{\beta}^{\text{EBLUE}})$ is the unshrunken predictor of the random effect for area *i*. It is not difficult to see that then $\hat{\mu}_j = \sum_{k \in s} \phi_{kj} y_k$ where $\phi_{kj} = c_{ijsk} + b_{ijsk} I(k \in i)$, with

$$\boldsymbol{c}_{ijs} = (c_{ijsk};\, k \in s)$$

$$= \hat{\Sigma}_{ss}^{-1} \boldsymbol{X}_s (\boldsymbol{X}_s^T \hat{\Sigma}_{ss}^{-1} \boldsymbol{X}_s)^{-1} \{\boldsymbol{x}_j - \boldsymbol{X}_{is}^T \boldsymbol{Z}_{is} (\boldsymbol{Z}_{is}^T \boldsymbol{Z}_{is})^{-1} \boldsymbol{z}_j\}$$

and $\boldsymbol{b}_{ijs} = (b_{ijsk};\, k \in s_i) = \boldsymbol{Z}_{is} (\boldsymbol{Z}_{is}^T \boldsymbol{Z}_{is})^{-1} \boldsymbol{z}_j$. Note that these ϕ_{kj}'s are also used to calculate the value of $\hat{\lambda}_j$ defined immediately after (7).

Finally, we observe that when (14) is used in (8), the estimated bias (9) becomes

$$\hat{B}(\hat{m}_i) = \sum_{h=1}^{D} \left(\sum_{j \in s_h} \boldsymbol{w}_{ij}^{\text{EBLUP}} \boldsymbol{z}_j \right)^T \tilde{\boldsymbol{u}}_h - \bar{\boldsymbol{z}}_i^T \tilde{\boldsymbol{u}}_i$$

since the EBLUP weights (13) are 'locally calibrated' on *X*, *i.e.*, $\sum_{j \in s} w_{ij}^{\text{EBLUP}} \boldsymbol{x}_j = \bar{\boldsymbol{x}}_i$. It follows that in this case the variable δ_i defined immediately before (11) takes the form

$$\delta_i = \sum_{h=1}^{D} w_{hi}^{\text{EBLUP}} \bar{\boldsymbol{z}}_h^T \tilde{\boldsymbol{u}}_h - \bar{\boldsymbol{z}}_i^T \tilde{\boldsymbol{u}}_i$$

where $w_{hi}^{\text{EBLUP}} = \sum_{j \in s_h} w_{ij}^{\text{EBLUP}}$. For a large enough overall sample size δ_i can be approximated by

$$\delta_i \approx \sum_{h=1}^{D} w_{hi}^{\text{BLUP}} \bar{\boldsymbol{z}}_h^T (\boldsymbol{Z}_{hs}^T \boldsymbol{Z}_{hs})^{-1} \boldsymbol{Z}_{hs}^T (y_{hs} - \boldsymbol{X}_{hs} \beta)$$

$$- \bar{\boldsymbol{z}}_i^T (\boldsymbol{Z}_{is}^T \boldsymbol{Z}_{is})^{-1} \boldsymbol{Z}_{is}^T (y_{is} - \boldsymbol{X}_{is} \beta)$$

$$= \sum_{\substack{h=1 \\ h \neq i}}^{D} w_{hi}^{\text{BLUP}} \bar{\boldsymbol{z}}_h^T \{\boldsymbol{u}_h + (\boldsymbol{Z}_{hs}^T \boldsymbol{Z}_{hs})^{-1} \boldsymbol{Z}_{hs}^T \boldsymbol{e}_{hs}\}$$

where w_{hi}^{BLUP} is the BLUP equivalent of w_{hi}^{EBLUP}. The variance of δ_i can therefore be estimated via

$$\hat{V}(\delta_i) = \sum_{\substack{h=1 \\ h \neq i}}^{D} (w_{hi}^{\text{EBLUP}})^2 \bar{\boldsymbol{z}}_h^T \{\hat{\Omega} + \hat{\sigma}^2 (\boldsymbol{Z}_{hs}^T \boldsymbol{Z}_{hs})^{-1}\} \bar{\boldsymbol{z}}_h. \tag{15}$$

If $\hat{V}(\delta_i)$ is small relative to the value of (7) in this case, then (8) can be used to estimate the MSE of the EBLUP. However, when n_i is very small, this condition may not hold. In such cases it may be advisable to consider more model-dependent MSE estimators like the Prasad-Rao (PR) MSE estimator (Prasad and Rao 1990; Rao 2003, section 7.2.3). When a random means model is assumed, but the between area variability is very small relative to the within area variability, this advice extends to moderate area sample sizes as we now show.

2.2.2 MSE estimation for the EBLUP under the random means model

The random means model is the special case of (12) where $y_j = \beta + u_i + e_j$, with $u_i \sim N(0, \sigma_u^2)$ and $e_j \sim N(0, \sigma^2)$. The EBLUE of β is then $\hat{\beta} = \sum_{h=1}^{D} \hat{\alpha}_h \bar{y}_{hs}$ with $\hat{\alpha}_i = (\hat{\varphi} + n_i^{-1})^{-1} \{\sum_{h=1}^{D} (\hat{\varphi} + n_h^{-1})^{-1}\}^{-1}$ and $\hat{\varphi} = \hat{\sigma}_u^2 / \hat{\sigma}^2$, and the EBLUP (13) is defined by weights of the form

$$w_{ij}^{\text{EBLUP}} = (1 - f_i)(1 - \hat{\gamma}_i) \sum_{h=1}^{D} \hat{\alpha}_h n_h^{-1} I(j \in h)$$

$$+ \{ f_i + (1 - f_i) \hat{\gamma}_i \} n_i^{-1} I(j \in i)$$

with $\hat{\gamma}_i = n_i \hat{\varphi} (1 + n_i \hat{\varphi})^{-1}$. For $j \in h$, $\hat{\mu}_j = \sum_{k \in s} \phi_{kj} y_k = \bar{y}_{hs}$ and so

$$\hat{\lambda}_j = (1 - \phi_{jj})^2 + \sum_{k \in s(-j)} \phi_{kj}^2$$

$$= (1 - n_h^{-1})^2 + (n_h - 1) n_h^{-2} = (n_h - 1) n_h^{-1}.$$

It follows that the estimator (7) of the conditional prediction variance of $\hat{m}_i^{\text{EBLUP}}$ in this case is

$$\hat{V}(\hat{m}_i^{\text{EBLUP}}) = (1 - f_i)^2 \Big[\sum_{h=1}^{D} \{ (1 - \hat{\gamma}_i)^2 \hat{\alpha}_h^2 n_h^{-2}$$

$$+ (N_i - n_i)^{-1} n^{-1} \} n_h s_h^2$$

$$+ \hat{\gamma}_i n_i^{-1} \{ 2 (1 - \hat{\gamma}_i) \hat{\alpha}_i + \hat{\gamma}_i \} s_i^2 \Big],$$

where $s_h^2 = (n_h - 1)^{-1} \sum_{j \in s_h} (y_j - \bar{y}_{hs})^2$, while from (9) the estimator of the conditional prediction bias of $\hat{m}_i^{\text{EBLUP}}$ is $\hat{B}(\hat{m}_i^{\text{EBLUP}}) = (1 - f_i)(1 - \hat{\gamma}_i)(\hat{\beta} - \bar{y}_{is})$. For $h \neq i$ we also then have

$$w_{hi}^{\text{EBLUP}} = \sum_{j \in s_h} w_{ij}^{\text{EBLUP}}$$

$$= (1 - f_i) \hat{\alpha}_h (1 + n_i \hat{\varphi})^{-1} \approx \hat{\alpha}_h (1 + n_i \hat{\varphi})^{-1}$$

when we ignore $O(N_i^{-1})$ terms. A similar approximation to (15) therefore leads to

$$\hat{V}(\delta_i) = \sum_{\substack{h=1 \\ h \neq i}}^{D} (w_{hi}^{\text{EBLUP}})^2 (\hat{\sigma}_u^2 + n_h^{-1} \hat{\sigma}^2)$$

$$\approx \hat{\sigma}^2 \sum_{\substack{h=1 \\ h \neq i}}^{D} \left(\frac{\hat{\alpha}_h}{1 + n_i \hat{\varphi}} \right)^2 \left(\frac{1 + n_h \hat{\varphi}}{n_h} \right).$$

Suppose now that the sample size in every small area is the same, *i.e.*, $n_i = m$. Then $n = mD$, $\hat{\alpha}_h = D^{-1}$ and the approximation to $\hat{V}(\delta_i)$ above takes the form

$$\hat{V}(\delta_i) = \hat{\sigma}^2 \sum_{\substack{h=1 \\ h \neq i}}^{D} \left(\frac{D^{-1}}{1 + m\hat{\varphi}} \right)^2 \left(\frac{1 + m\hat{\varphi}}{m} \right) \approx n^{-1} (1 + m\hat{\varphi})^{-1} \hat{\sigma}^2$$

while the corresponding approximation to $\hat{V}(\hat{m}_i^{\text{EBLUP}})$ is

$$\hat{V}(\hat{m}_i^{\text{EBLUP}}) \approx \sum_{h=1}^{D} (1 + m\hat{\varphi})^{-2} D^{-2} m^{-1} s_h^2$$

$$+ (1 + m\hat{\varphi})^{-2} \hat{\varphi} (2D^{-1} + m\hat{\varphi}) s_i^2$$

$$= n^{-1} (1 + m\hat{\varphi})^{-2} \left\{ \left(D^{-1} \sum_{h=1}^{D} s_h^2 \right) + m\hat{\varphi} (2 + n\hat{\varphi}) s_i^2 \right\}.$$

Comparing these approximations to $\hat{V}(\delta_i)$ and $\hat{V}(\hat{m}_i^{\text{EBLUP}})$ we see that if $m\hat{\varphi}$ is small (*e.g.*, when m and $\hat{\varphi}$ are both small) then these terms will be of similar magnitude. In this situation we expect (8) to overestimate the true MSE of the EBLUP. In particular, the approximation to (8) when $m\hat{\varphi}$ is small and N_i is large is

$$\hat{M}(\hat{m}_i^{\text{EBLUP}}) \approx n^{-1} \left(D^{-1} \sum_{h=1}^{D} s_h^2 \right) + \left(\bar{y}_{is} - D^{-1} \sum_{h=1}^{D} \bar{y}_{hs} \right)^2. \quad (16)$$

Note that the expectation of the squared residual on the right hand side of (16) when $m\hat{\varphi}$ is small is $(1 - D^{-1})(\sigma_u^2 + m^{-1}\sigma^2) = O(1)$ and so it is the leading term in this estimator in this situation. This expression can be compared with the corresponding one for the MSE estimator of the EBLUP suggested by Prasad and Rao (1990). Under the random means model, the PR MSE estimator is

$$\hat{M}_{\text{PR}}(\hat{m}_i^{\text{EBLUP}}) = (1 - f_i)^2 \Big[\hat{\gamma}_i m^{-1} \hat{\sigma}^2$$

$$+ (1 - \hat{\gamma}_i)^2 \left(m \sum_{h=1}^{D} \hat{\tau}_h^{-1} \right)^{-1} + N_i^{-1} (1 - f_i)^{-1} \hat{\sigma}^2$$

$$+ \frac{2}{T} m \hat{\tau}_i^{-3} \left\{ \hat{\sigma}^4 \left(\frac{n - D}{\hat{\sigma}^4} + \sum_{h=1}^{D} \hat{\tau}_h^{-2} \right) \right.$$

$$\left. + \hat{\sigma}_u^4 m^2 \sum_{h=1}^{D} \hat{\tau}_h^{-2} + 2 \hat{\sigma}^2 \hat{\sigma}_u^2 m \sum_{h=1}^{D} \hat{\tau}_h^{-2} \right\} \Big]$$

where $\hat{\tau}_i = n_i \hat{\sigma}_u^2 + \hat{\sigma}^2$ and

$$T = \frac{n - D}{\hat{\sigma}^4} \sum_{h=1}^{D} n_h^2 \hat{\tau}_h^{-2}$$

$$+ \left(\sum_{h=1}^{D} \hat{\tau}_h^{-2} \right) \left(\sum_{h=1}^{D} n_h^2 \hat{\tau}_h^{-2} \right) - \left(\sum_{h=1}^{D} n_h \hat{\tau}_h^{-2} \right)^2.$$

Assuming $n_i = m$, $m\hat{\varphi}$ is small and N_i is large, $\hat{M}_{\text{PR}}(\hat{m}_i^{\text{EBLUP}})$ has the approximation

$$\hat{M}_{\text{PR}}(\hat{m}_i^{\text{EBLUP}}) \approx \hat{\sigma}^2\{n^{-1} + 2(n-D)^{-1}\} + \hat{\sigma}_u^2. \qquad (17)$$

Comparing (16) and (17) we can see that the instability and the overestimation associated with the use of (8) in this situation are both due to the use of the square of the single degree of freedom area level residual $\bar{y}_{is} - D^{-1}\sum_{h=1}^{D}\bar{y}_{hs}$ as an estimator of σ_u^2. This reinforces earlier comments that (8) should not generally be used for estimating the MSE of the EBLUP if the area sample sizes are very small or, in the special case of the random means model, for moderate area sample sizes when the between area variability is very small relative to the within area variability.

2.2.3 MSE estimation for the MBDE

The second predictor of m_i that we consider is the Model-Based Direct Estimator (MBDE) described in Chandra and Chambers (2009). This is based on the same linear mixed model (12) as the EBLUP, with the MBDE predictor defined as

$$\hat{m}_i^{\text{MBDE}} = \sum\nolimits_{j \in s} w_{ij}^{\text{MBDE}} y_j = (\boldsymbol{w}_{is}^{\text{MBDE}})^T \boldsymbol{y}_s \qquad (18)$$

where

$$w_{ij}^{\text{MBDE}} = \frac{I(j \in s_i)\, w_j^{\text{EBLUP}}}{\sum_{k \in s} I(k \in s_i)\, w_k^{\text{EBLUP}}}. \qquad (19)$$

Here $I(j \in s_i)$ is the indicator function for unit j to be in the area i sample, and $\boldsymbol{w}_s^{\text{EBLUP}} = (w_j^{\text{EBLUP}})$ is the vector of weights that defines the EBLUP for the population total of the y_j under (12), *i.e.*,

$$\boldsymbol{w}_s^{\text{EBLUP}} = (w_j^{\text{EBLUP}}) = \mathbf{1}_n + \{\hat{\boldsymbol{H}}_s^T \boldsymbol{X}_r^T + (\mathbf{I}_n - \hat{\boldsymbol{H}}_s^T \boldsymbol{X}_s^T)\hat{\boldsymbol{\Sigma}}_{ss}^{-1}\hat{\boldsymbol{\Sigma}}_{sr}\}\mathbf{1}_{N-n}$$

where $\mathbf{1}_n$ $(\mathbf{1}_{N-n})$ denotes the unit vector of size n $(N - n)$ and $\hat{\boldsymbol{H}}_s$ was defined in section 2.2.1. In this case pseudo-linearisation based estimation of the area-specific MSE of the MBDE is carried out using (8), with weights defined by (19). Note that the estimated expected values used in (8) when applied to the MBDE are the same as the unshrunken estimates (14) used with the EBLUP, reflecting the fact that both the MBDE and the EBLUP are based on the same linear mixed model (12). However, the MBDE weights (19) are not locally calibrated, and so the squared bias term in (8) cannot be ignored when estimating the MSE of this predictor. Furthermore, since

$$w_{hi}^{\text{MBDE}} = \sum\nolimits_{j \in s_h} w_{ij}^{\text{MBDE}} = 0$$

for $h \neq i$, we have $\delta_i = 0$ for the MBDE and so the bias estimator (9) works well in this case.

2.2.4 MSE estimation for the M-quantile estimator

The third estimator that we consider is based on the M-quantile modelling approach described in Chambers and Tzavidis (2006). This approach does not assume an underlying linear mixed model, relying instead on characterising the relationship between y_j and $\boldsymbol{x}_j$ in area i in terms of the linear M-quantile model that best 'fits' the sample y_j values from this area. That is, this approach replaces (12) by a model of the form

$$\boldsymbol{y}_i = \boldsymbol{X}_i \beta(q_i) + \boldsymbol{e}_i \qquad (20)$$

where $\beta(q)$ denotes the coefficient vector of a linear model for the regression M-quantile of order q for the population values of Y and X, and q_i denotes the M-quantile coefficient of area i. Given an estimate $\hat{q}_i$ of q_i, an iteratively re-weighted least squares (IRLS) algorithm is used to calculate an estimate

$$\hat{\beta}(\hat{q}_i) = \left\{\boldsymbol{X}_s' \boldsymbol{W}_s(\hat{q}_i)\boldsymbol{X}_s\right\}^{-1} \boldsymbol{X}_s' \boldsymbol{W}_s(\hat{q}_i)\boldsymbol{y}_s \qquad (21)$$

of $\beta(q_i)$ in (20), and a non-sample value of y_j in area i is then predicted by $\hat{y}_j = \boldsymbol{x}_j^T \hat{\beta}(\hat{q}_i)$. Here $\boldsymbol{W}_s(\hat{q}_i)$ is the diagonal matrix of final weights used in the IRLS algorithm.

Tzavidis, Marchetti and Chambers (2010) note that value of the M-quantile estimator suggested in Chambers and Tzavidis (2006) can be interpreted as the expected value of Y in area i with respect to a biased estimator of the distribution function of this variable in the area. They therefore develop an improved M-quantile estimator, replacing this biased distribution function estimator by the Chambers and Dunstan (1986) distribution function estimator under the area-specific model (1). This corresponds to predicting m_i by

$$\hat{m}_i^{\text{MQ}} = \sum\nolimits_{j \in s} w_{ij}^{\text{MQ}} y_j = (\boldsymbol{w}_{is}^{\text{MQ}})^T \boldsymbol{y}_s \qquad (22)$$

where

$$\boldsymbol{w}_{is}^{\text{MQ}} = n_i^{-1}\boldsymbol{\Delta}_{is} + (1 - N_i^{-1}n_i)\boldsymbol{W}_s(\hat{q}_i)\boldsymbol{X}_s\{\boldsymbol{X}_s^T \boldsymbol{W}_s(\hat{q}_i)\boldsymbol{X}_s\}^{-1}(\bar{\boldsymbol{x}}_{ir} - \bar{\boldsymbol{x}}_{is}).$$

Here $\bar{\boldsymbol{x}}_{is}$ and $\bar{\boldsymbol{x}}_{ir}$ are the vectors of sample and non-sample means of the x_j in area i. It is not difficult to show that the weights following (22) are locally calibrated. Furthermore, if we then put $\hat{\mu}_j = \boldsymbol{x}_j^T \hat{\beta}(\hat{q}_i)$, where $\hat{\beta}(\hat{q}_i)$ is defined by (21), it is easy to see that (9) is zero and so the area-specific MSE of the bias-corrected M-quantile estimator (22) can be estimated using just the estimated prediction variance component (7). Since the constant $\hat{\lambda}_j$ in (7) is typically very close to one under M-quantile estimation, we set it equal to this value whenever we compute values of (7) that relate to small area estimation (SAE) under the M-quantile modelling approach.

As we have already done with the EBLUP, we note that use of (7) implicitly treats the weights defining (22) as fixed, which is actually not the case since the matrix $\boldsymbol{W}_s(\hat{q}_i)$ is a function of the sample values of Y. An immediate consequence is that pseudo-linearisation based estimation of the MSE of the M-quantile predictor via (7) is a first order approximation to the true MSE of this estimator. Nevertheless, since accounting for weight variability in the definition of the M-quantile estimator considerably complicates estimation of its MSE - see Street, Carroll and Ruppert (1988) for an examination of this issue in the context of 'standard' M-estimation of regression coefficients - it is of interest to see how the relatively simple estimator (7) performs when used to estimate this MSE.

2.3 MSE estimation for the pseudo-linear synthetic EBLUP

In many SAE applications there are areas that contain no sample, and hence synthetic estimation is used. Although such estimators do not fit into the class of pseudo-linear estimators considered in this paper, the ideas behind the conditional MSE estimator (8) can be applied here as well. To see this, assume that these areas are numbered last, *i.e.*, if D^+ areas have non-zero sample then $n_h > 0$ for $h \le D^+$ and $n_h = 0$ for $h > D^+$. For $i > D^+$ the 'synthetic EBLUP' for m_i is

$$\begin{aligned}\hat{m}_i^{\text{SYN-EBLUP}} &= \bar{\boldsymbol{x}}_i^T \hat{\beta}^{\text{EBLUE}} = (\boldsymbol{w}_{is}^{\text{SYN-EBLUP}})^T \boldsymbol{y}_s \\ &= \sum\nolimits_{h=1}^{D^+} \sum\nolimits_{j \in s_h} w_{ij}^{\text{SYN-EBLUP}} y_j \qquad (23)\end{aligned}$$

where

$$\boldsymbol{w}_{is}^{\text{SYN-EBLUP}} = (w_{ij}^{\text{SYN-EBLUP}}) = \hat{\boldsymbol{H}}_s^T \bar{\boldsymbol{x}}_i .$$

Clearly (23) is a pseudo-linear estimator, and so we can use (7) to estimate its prediction variance, observing that since $n_i = 0$, $a_{ij} = N_i w_{ij}^{\text{EBLUP}}$ and so (7) becomes

$$\hat{V}(\hat{m}_i^{\text{SYN-EBLUP}}) = \sum\nolimits_{j \in s} \{(w_{ij}^{\text{SYN-EBLUP}})^2 + N_i^{-1} n^{-1}\} \hat{\lambda}_j^{-1} (y_j - \hat{\mu}_j)^2 . \quad (24)$$

Unfortunately, since there is no sample in area i, we cannot use (9) to estimate the area-specific bias (2) of $\hat{m}_i^{\text{SYN-EBLUP}}$. However, under the linear mixed model (12), this bias has expected value

$$E(\hat{m}_i^{\text{SYN-EBLUP}} - m_i) = \sum\nolimits_{h=1}^{D^+} \sum\nolimits_{j \in s_h} w_{ij}^{\text{SYN-EBLUP}} \left(\boldsymbol{x}_j^T \beta + \boldsymbol{z}_j^T \boldsymbol{u}_h \right) - \bar{\boldsymbol{x}}_i^T \beta - \bar{\boldsymbol{z}}_i^T \boldsymbol{u}_i .$$

The conditional expectation of the square of this expected bias, given the area effects $\boldsymbol{u}_s = (u_h;\, h = 1, \ldots, D^+)$ for the sampled areas, is

$$E\{E^2(\hat{m}_i^{\text{SYN-EBLUP}} - m_i) \,|\, \boldsymbol{X}, \boldsymbol{u}_s\} = \left\{ \sum\nolimits_{h=1}^{D^+} \sum\nolimits_{j \in s_h} w_{ij}^{\text{SYN-EBLUP}} (\boldsymbol{x}_j^T \beta + \boldsymbol{z}_j^T \boldsymbol{u}_h) - \bar{\boldsymbol{x}}_i^T \beta \right\}^2 + \bar{\boldsymbol{z}}_i^T \Omega \bar{\boldsymbol{z}}_i ,$$

which immediately suggests that for a non-sampled area i we estimate the squared bias of the synthetic estimator $\hat{m}_i^{\text{SYN-EBLUP}}$ by

$$\hat{B}^2(\hat{m}_i^{\text{SYN-EBLUP}}) = \left\{ \sum\nolimits_{h=1}^{D^+} \sum\nolimits_{j \in s_h} w_{ij}^{\text{SYN-EBLUP}} \left(\boldsymbol{x}_j^T \hat{\beta}^{\text{EBLUE}} + \boldsymbol{z}_j^T \tilde{\boldsymbol{u}}_h \right) - \bar{\boldsymbol{x}}_i^T \hat{\beta}^{\text{EBLUE}} \right\}^2 + \bar{\boldsymbol{z}}_i^T \hat{\Omega} \bar{\boldsymbol{z}}_i . \quad (25)$$

Here $\tilde{\boldsymbol{u}}_h$ is the 'unshrunken' estimated effect for sampled area h – see (14). Our proposed MSE estimator for $\hat{m}_i^{\text{SYN-EBLUP}}$ is then the sum of (24) and (25). Note that, unlike (8), this MSE estimator includes no information from area i, and so is not an estimator of the area-specific MSE of (23). In particular, its validity depends completely on the mixed model (12) holding, and so it is not robust to misspecification of this model.

3. Simulation studies of the proposed MSE estimator

In this section we describe results from five simulation studies that aim at assessing the performance of the approach to conditional MSE estimation described in the previous section. Three of these studies are model-based simulations, with population data generated from the linear mixed model (12). The remaining two are design-based simulations, with population data derived from two real survey datasets where linear SAE is of interest.

Given our focus on bias-robustness, the main performance indicator for an MSE estimator in all five studies is its median relative bias, defined by

$$\text{RB}(M) = \underset{i}{\text{median}} \left\{ M_i^{-1} K^{-1} \sum\nolimits_{k=1}^{K} (\hat{M}_{ik} - M_i) \right\} \times 100.$$

Here the subscript i indexes the small areas and the subscript k indexes the K Monte Carlo simulations, with $\hat{M}_{ik}$ denoting the simulation k value of the MSE estimator in area i, and M_i denotes the actual (*i.e.*, Monte Carlo) MSE in area i. Since we would naturally prefer to use the more stable of two approximately unbiased MSE estimators, we also measured the stability of an MSE estimator by its median percent relative root mean squared error,

$$\text{RRMSE}(M) = \underset{i}{\text{median}}\left\{\sqrt{K^{-1}\sum\nolimits_{k=1}^{K}\left(\frac{\hat{M}_{ik}-M_i}{M_i}\right)^2}\right\}\times 100.$$

Although the purpose of this paper is not to compare different methods of SAE, it is useful to relate MSE estimation performance for a particular method of SAE to the actual estimation performance of this method. We therefore provide two measures of the relative performance of the SAE methods that were used in our simulations. These are the median percent relative bias

$$\text{RB}(m) = \underset{i}{\text{median}}\left\{\bar{m}_i^{-1}K^{-1}\sum\nolimits_{k=1}^{K}\left(\hat{m}_{ik}-m_{ik}\right)\right\}\times 100$$

and the median percent relative root mean squared error

$$\text{RRMSE}(m) = \underset{i}{\text{median}}\left\{\sqrt{K^{-1}\sum\nolimits_{k=1}^{K}\left(\frac{\hat{m}_{ik}-m_{ik}}{m_{ik}}\right)^2}\right\}\times 100$$

of the estimates $\hat{m}_{ik}$ generated by an estimation method. Note that $\bar{m}_i = K^{-1}\sum_{k=1}^{K} m_{ik}$ here.

3.1 Model-based simulations

The first model-based simulation study was based on population data generated under the mixed model (12) with Gaussian random effects. It used a population size of $N =$ 15,000, with $D = 30$ small areas. Population sizes in the small areas were uniformly distributed over the interval [443, 542] and kept fixed over simulations. At each simulation, population values for Y were generated under the random intercepts model $y_j = 500 + 1.5x_j + u_i + e_j$, with x_j drawn from a chi-squared distribution with 20 degrees of freedom. The area effects u_i and individual effects e_j were independently drawn from $N(0, \sigma_u^2)$ and $N(0, \sigma^2)$ distributions respectively, with the values of σ_u and σ shown in rows SIM1-A and SIM1-B of Table 1. A sample of size $n = 600$ was selected from each simulated population, with area sample sizes proportional to the fixed area populations, resulting in a median area sample size of $n_i = 20$. Sampling was via stratified random sampling, with the strata defined by the small areas. A total of $K = 1{,}000$ simulations were carried out.

Conditions for the second model-based simulation study were the same as in the first, with the exception that the area level random effects and the individual level random effects were independently drawn from mean corrected chi-square distributions respectively. The corresponding values of the area level and individual level variances are shown in rows SIM2-A and SIM2-B in Table 1. Finally, in the third model-based simulation study conditions were kept the same as in SIM1-A and SIM1-B for areas 1-25, but in areas 26-30 the area effects were independently drawn from a normal distribution with a larger variance. We refer to this as a Mixture in Table 1, with variances for areas 1-25 shown in rows SIM3-A and SIM3-B, and variances for areas 26-30 shown in rows SIM3-A* and SIM3B*. Our objective in this third simulation was to investigate the behaviour of the different methods of MSE estimation for 'outlier' areas, and so we show values relating to areas 1-25 and 26-30 separately in Tables 2 and 4. We also replicated all three scenarios above using a reduced overall sample size of $n = 150$ (with median area sample size $n_i = 5$). These additional simulations allowed us to investigate the effect of reduced sample sizes on the performance of the MSE estimators.

Table 1
Parameter values used in model-based simulations

Type	Simulation	σ_u^2	σ^2	$\rho = \sigma_u^2(\sigma_u^2+\sigma^2)^{-1}$
Gaussian	SIM1-A	10.40	94.09	0.1
	SIM1-B	40.32	94.09	0.3
Chi-square	SIM2-A	2.0	10.0	0.1667
	SIM2-B	4.0	10.0	0.2857
Mixture (areas 1-25)	SIM3-A	10.40	94.09	0.10
	SIM3-B	40.32	94.09	0.30
Mixture (areas 26-30)	SIM3-A*	225.0	94.09	0.7051
	SIM3-B*	225.0	94.09	0.7051

Table 2
Median relative biases RB(*m*) and median relative root mean squared errors RRMSE(*m*) of estimators of small area means in model-based simulations

Weighting Method	Simulation							
	SIM1-A	**SIM1-B**	**SIM2-A**	**SIM2-B**	**SIM3-A**	**SIM3-B**	**SIM3-A***	**SIM3-B***
RB(*m*), median $n_i = 20$								
Regression	0.005	0.005	0.000	0.000	0.004	0.004	0.006	0.006
EBLUP, (13)	0.005	0.006	0.004	-0.002	0.004	0.005	0.006	0.005
MBDE, (18)	0.006	0.006	0.005	-0.008	0.007	0.007	0.001	0.001
M-quantile, (22)	0.009	0.008	-0.002	0.002	0.015	0.015	-0.013	-0.013
RRMSE(*m*), median $n_i = 20$								
Regression	0.40	0.40	0.13	0.13	0.40	0.40	0.41	0.41
EBLUP, (13)	0.35	0.38	0.12	0.13	0.37	0.38	0.45	0.42
MBDE, (18)	0.55	0.55	0.41	0.43	0.56	0.56	0.55	0.55
M-quantile, (22)	0.41	0.41	0.13	0.13	0.41	0.41	0.36	0.36
RB(*m*), median $n_i = 5$								
Regression	-0.002	-0.003	-0.001	0.002	-0.003	-0.004	0.011	0.011
EBLUP, (13)	0.001	0.005	-0.002	0.003	0.002	-0.001	0.008	0.011
MBDE, (18)	-0.002	-0.002	-0.005	0.004	-0.001	-0.002	-0.002	-0.002
M-quantile, (22)	-0.001	-0.001	-0.001	0.001	-0.003	-0.003	0.014	0.014
RRMSE(*m*), median $n_i = 5$								
Regression	0.81	0.81	0.26	0.26	0.82	0.82	0.80	0.80
EBLUP, (13)	0.53	0.69	0.19	0.22	0.61	0.71	1.00	0.87
MBDE, (18)	1.13	1.13	0.83	0.83	1.13	1.13	1.13	1.13
M-quantile, (22)	0.81	0.81	0.26	0.26	0.81	0.81	0.80	0.80

Table 2 shows the median bias RB(*m*) and median relative root mean squared error RRMSE(*m*) of the SAE methods investigated in our simulations for the two sample sizes ($n = 600$ and 150). These are the synthetic regression estimator (see Rao 2003, page 136), the EBLUP with weights defined by (13), the MBDE with weights defined by (18) and the M-quantile estimator defined by the weights (22). The differences between the various SAE estimators in Table 2 are essentially as one would expect. Bias is not really an issue (to be expected given the population data follow a linear model in all cases), while for Simulation scenarios 1 and 2 the indirect estimator (EBLUP) is the most efficient in terms of RRMSE. The M-quantile estimator is the best performer for SIM3-A* and SIM3-B* with $n_i = 20$ but its difference from the regression synthetic estimator reduces for the scenario with the smaller area-specific sample sizes. Note that in this case the M-quantile weights (22) are based on an outlier-robust estimate of the M-quantile coefficient $\hat{q}_i$ for area *i*, defined by the median (rather than the mean) of the M-quantile coefficients of sampled units in this area. Further, as the sample sizes decrease, the RRMSEs of all estimators increase, but their relative performances remain the same. Under normality the EBLUP is better than the M-quantile estimator but the differences between these two estimators become smaller as we move away from normality, with the M-quantile estimator more efficient in the mixture model scenarios.

Table 3 sets out the various MSE estimators investigated in our simulations that are based on the approach proposed in this paper. These are collectively referred to as “conditional” MSE estimators below. In Table 4 we show the performances of MSE estimators for the small area estimators considered in Table 2. Note that in addition to the conditional MSE estimators, we provide results for three other MSE estimators for the EBLUP, with PR0 denoting the estimator suggested by Prasad and Rao (1990), see Rao (2003, section 6.2.6). It is noteworthy that PR0 is not an estimator of the area-specific MSE of the EBLUP, but of its MSE under the mixed linear model (12), *i.e.*, averaged over possible realisations of the area effect. In contrast, the MSE estimators PR1 and PR2 in Table 4 are the area-specific versions of PR0 suggested in Rao (2003, section 6.3.2 expressions 6.3.15 and 6.3.16 respectively). Finally, we note that the MSE estimator of the synthetic regression estimator that we used in our simulations is its variance estimator based on a fixed effects population regression model. We denote it by VReg.

The results set out in Table 4 focus on the median biases RB(M) and median relative root mean squared error RRMSE(M) of the various MSE estimators. Not surprisingly, given that all its underlying assumptions are met, the PR0 estimator and its area-specific alternatives, PR1 and PR2, perform very well in both normal scenarios (SIM1-A and SIM1-B) and both chi-squared scenarios (SIM2-A and SIM2-B), with virtually no bias ($n_i = 20$) or small bias when within area sample sizes are very small. For the MSE estimator of the synthetic regression estimator, on the other hand, we see substantial relative bias under all simulation scenarios.

Table 3
Definitions of conditional MSE estimators for different weighting methods

Weighting Method	Definition of $\hat{\mu}_j, j \in i$	MSE Estimator
EBLUP (13)	(14)	(8)
MBDE (18)	(14)	(8)
M-quantile (22)	$x_j^T \hat{\beta}(\hat{q}_i)$	(7) with $\hat{\lambda}_j = 1$
Synthetic EBLUP (23)	(14)	(24) + (25)

Table 4
Median relative biases RB(*M*) and median relative root mean squared errors RRMSE(*M*) for MSE estimators in model-based simulations

Weighting	MSE	Simulation							
Method	Estimator	SIM1-A	SIM1-B	SIM2-A	SIM2-B	SIM3-A	SIM3-B	SIM3-A*	SIM3-B*
				RB(*M*), median $n_i = 20$					
Regression	VReg	7.59	21.82	11.81	20.78	23.66	34.27	23.97	34.64
EBLUP, (13)	PR0	-0.83	-0.72	0.56	1.16	3.44	0.71	-15.65	-6.51
	PR1	-0.97	-0.72	0.64	1.08	2.94	0.56	-13.70	-5.81
	PR2	-0.92	-0.72	0.64	1.16	3.20	0.61	-14.65	-6.19
	Conditional	3.89	-0.89	3.06	0.93	-0.05	-0.54	-2.56	-1.59
MBDE, (18)	Conditional	-0.81	-0.80	-0.06	-0.42	-0.75	-0.75	-0.98	-0.98
M-quantile, (22)	Conditional	-3.10	-1.66	-0.09	-1.90	-5.04	-3.17	11.26	11.04
				RRMSE(*M*), median $n_i = 20$					
Regression	VReg	18	51	30	53	59	85	60	86
EBLUP, (13)	PR0	12	7	15	10	11	7	29	14
	PR1	14	7	17	11	10	7	27	13
	PR2	12	7	16	10	11	7	28	13
	Conditional	62	31	70	49	31	30	42	32
MBDE, (18)	Conditional	70	70	126	128	71	71	67	67
M-quantile, (22)	Conditional	32	34	49	48	31	32	48	48
				RB(*M*), median $n_i = 5$					
Regression	VReg	5.59	19.17	10.35	19.12	20.92	30.91	22.93	33.00
EBLUP, (13)	PR0	3.51	-0.20	2.42	1.19	12.79	3.86	-30.64	-15.92
	PR1	3.04	-0.50	2.13	1.00	10.84	3.10	-25.77	-13.62
	PR2	3.16	-0.31	2.31	1.11	11.81	3.48	-28.16	-14.77
	Conditional	37.52	4.38	24.11	8.93	8.18	1.50	-0.66	-0.68
MBDE, (18)	Conditional	-0.24	-0.21	0.02	-0.09	-0.62	-0.33	1.29	1.24
M-quantile, (22)	Conditional	-7.60	-6.17	5.70	5.00	-5.95	-5.60	5.89	3.60
				RRMSE(*M*), median $n_i = 5$					
Regression	VReg	17	46	33	51	54	78	59	83
EBLUP, (13)	PR0	31	14	33	22	36	16	53	31
	PR1	48	18	44	28	34	16	48	29
	PR2	36	15	36	24	34	15	50	29
	Conditional	234	81	193	121	86	66	86	70
MBDE, (18)	Conditional	79	79	133	129	79	79	83	83
M-quantile, (22)	Conditional	62	63	90	97	63	63	122	102

The conditional MSE estimator for the EBLUP shows positive bias under both the normal (SIM1A) and chi-squared (SIM2A) scenarios, particularly for moderate intra-cluster correlation (3.89% and 37.52% for the normal scenario with 20 and 5 units in each area respectively and 3.06% and 24.11% for the chi-squared scenario with 20 and 5 units in each area respectively). This bias increases with decreasing sample size. However, things change when we examine the results for the outlier components of the mixture model scenarios (SIM3-A* and SIM3-B*). Here we see a substantial negative bias for all three versions of PR (ranging from -30.64% to -5.81% depending on the area sample sizes). In comparison, the conditional MSE estimator for the EBLUP now shows a smaller negative bias (-2.56% and -0.66%) while the same MSE estimator applied to the M-quantile estimator shows an upward bias. The conditional MSE estimator for the MBDE is essentially unbiased. Given that as far as MSE estimation is concerned, positive bias is preferable to negative bias, it seems clear that the proposed conditional MSE estimator is better able to handle this outlier situation. Figure 1 graphically illustrates this point for sample size $n = 600$. Here we show the area-specific RMSEs and the average (over the simulations) of the estimated RMSEs in each of the 30 areas for the mixture simulations SIM3-A and SIM3-A*, with the vertical line delineating the five 'outlier' areas. In the top panel of this plot we see that the PR0 estimator is unable to detect the step increase in the MSE of the EBLUP for these 'outlier' areas, being biased slightly high in the 'well-behaved' areas and then biased rather low in the 'outlier' areas. In contrast, the conditional MSE estimator for the EBLUP and the MBDE tracks the area specific RMSEs rather well, while the same MSE estimator based on M-quantile weights tends to be biased low in the 'well-behaved' areas, and biased high in the 'outlier' areas, which can be argued as being perhaps a rather better outcome than that recorded by the PR0 estimator in this simulation. It should be noted here that in certain circumstances an assumed model can be revised after outlier detection. However, this requires a sufficiently large number of detected outliers to permit their separate modelling. This is unlikely to happen in practice. Also, particular care must be taken with extrapolation of these results to the case of very small area sample sizes because of the instability that the conditional MSE estimator can exhibit in this case.

Table 4 also shows the relative RMSEs of the different MSE estimators across the three types of model-based simulation. Here we see that all three versions of the PR estimator of the MSE of the EBLUP are more stable than the conditional MSE estimator of the EBLUP (12% for PR *vs*. 62% for the conditional MSE for SIM1-A with $n_i = 20$ and 31% for PR *vs*. 234% for the conditional MSE for SIM1-A with $n_i = 5$). These differences decrease under scenarios SIM3-A* and SIM3-B*, however, although the PR MSE estimator remains more stable (13% for PR *vs*. 32% for the conditional MSE estimator for SIM3-B* with $n_i = 20$ and 29% for the PR MSE estimator *vs*. 70% for the conditional MSE estimator for SIM3-B* with $n_i = 5$). The same is true for the conditional MSE estimators of the MBDE and the M-quantile estimators. Essentially, given sample data that follow a mixed linear model, the PR MSE estimator of MSE is very stable, while the conditional MSE estimator is more variable.

In summary, although all methods of MSE estimation that we evaluated exhibited some bias for very small area sample sizes, our model-based simulation results provide evidence that for larger area sample sizes the conditional MSE estimation method (8) is bias robust when applied to the three pseudo-linear small area estimators EBLUP, MBDE and M-quantile. For very small area sample sizes its bias robustness is less evident. As one might expect, the model dependent 'area-averaged' MSE estimator PR0 for the EBLUP exhibits bias under model failure. The fact that we observed rather similar behaviour for the area-specific versions PR1 and PR2 of this MSE estimator indicates that 'area specific' does not necessarily mean 'bias robust'. In particular, the fact that PR1 and PR2 behave very similarly to PR0 may be because the area-specific components of PR1 and PR2 are of lower order and all three MSE estimators have the same leading term, which is not area-specific. Our results also show that the conditional MSE estimator (8) is much more variable than the model dependent PR MSE estimator, even for moderate area sample sizes.

3.2 Design-based simulations

What happens when, as in real life, we cannot be confident that our data follow a linear mixed model? In order to investigate this situation, we report results from two design-based simulation studies, both based on realistic populations, where a linear model assumption is essentially an approximation. The first involved a sample of 3,591 households spread across $D = 36$ districts of Albania that participated in the 2002 Albanian Living Standards Measurement Study. This sample was bootstrapped to create a realistic population of $N = 724{,}782$ households by re-sampling with replacement with probability proportional to a household's sample weight. A total of $K = 1{,}000$ independent stratified random samples were then drawn from this bootstrap population, with total sample size equal to that of the original sample and with districts defining the strata. Sample sizes within districts were the same as in the original sample, and varied between 8 and 688 (with median district sample size equal to 56). The *Y* variable of interest was household per capita consumption expenditure (HCE) and *X* was defined by three zero-one variables (ownership of television, parabolic antenna and land). The aim was to estimate the average value of HCE for each district. In the original 2002 survey, the linear relationship between HCE and the three variables making up *X* was rather weak, with very low predictive power. In particular, only ownership of land was significantly related to HCE at the five percent level. This fit was considerably improved by extending the linear model to include random intercepts, defined by independent district effects. These explained approximately 10 per cent of the residual variation in this model.

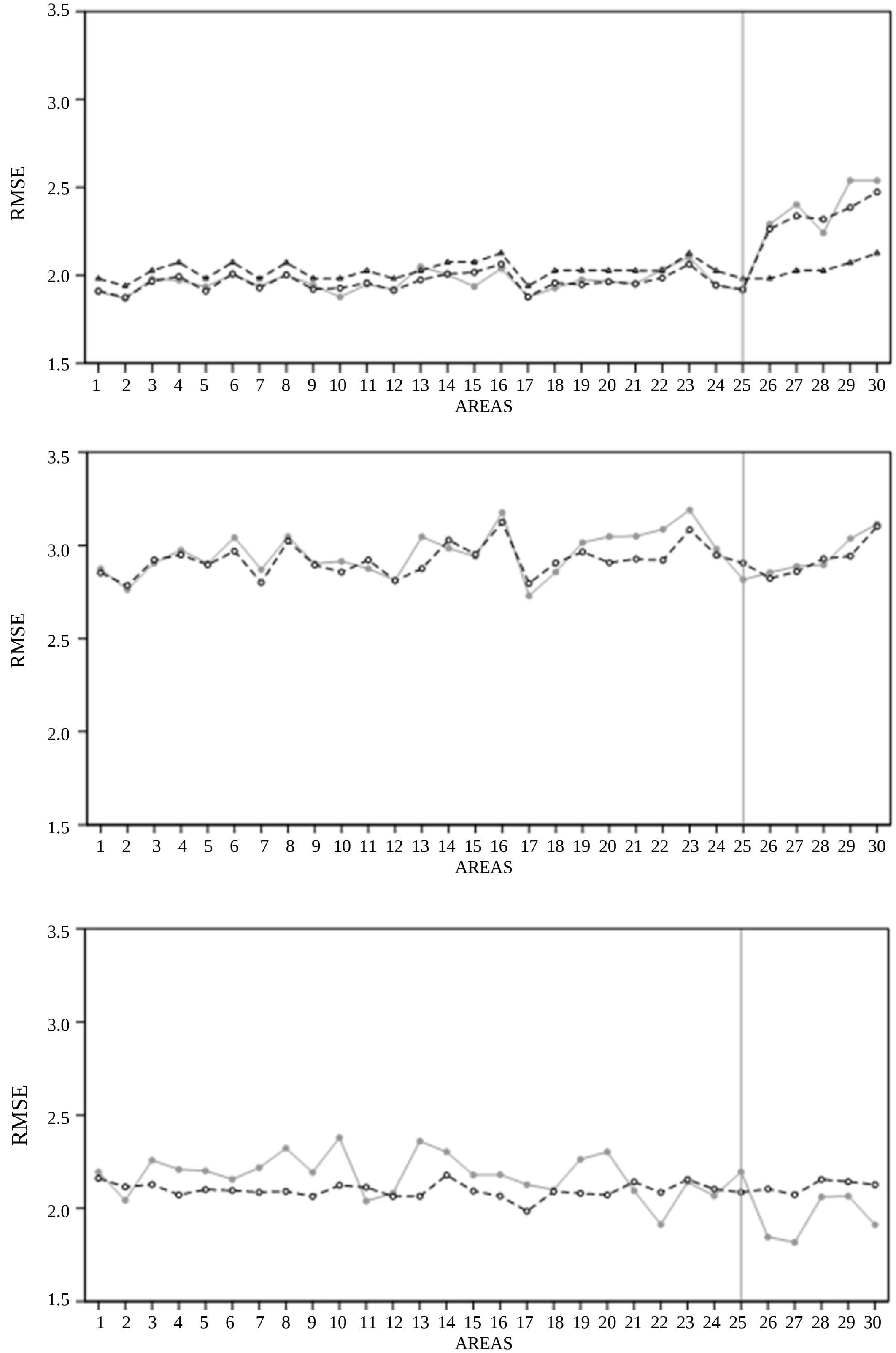


Figure 1 Area specific values of true RMSE (solid line) and average estimated RMSE (dashed line) obtained in the mixture-based simulations SIM3-A and SIM3-A*. Values for the PR0 estimator are indicated by Δ while those for the conditional estimator are indicated by o. Plots show results for the EBLUP (top), MBDE (centre) and M-quantile (bottom) estimators. Vertical line separates areas 26-30 with 'outlier' effects from 'well-behaved' areas 1-25. Total sample size is 600 with area-specific sample sizes equal to 20

The second design-based simulation study was based on an ‘outlier free’ version of the population of Australian broadacre farms that was used in the simulation studies reported in Chambers and Tzavidis (2006) and Chandra and Chambers (2009). In particular, this population was defined by bootstrapping a sub-sample of 1,579 ‘non-outlier’ farms that participated in the Australian Agricultural and Grazing Industries Survey (AAGIS) to create a population of $N =$ 78,072 farms by re-sampling from the original AAGIS sample with probability proportional to a farm’s sample weight. The small areas of interest in this case were the $D = 28$ broadacre farming regions represented in this sub-sample. The design-based simulation was carried out by selecting $K = 1{,}000$ independent stratified random samples from this bootstrap population, with strata defined by the regions and with stratum sample sizes defined by those in the original AAGIS sample. These sample sizes vary from 6 to 117, with a median region sample size of 53. Here Y is Total Cash Costs (TCC) associated with operation of the farm, and X is a vector that includes farm area (Area), effects for six post-strata defined by three climatic zones and two farm size bands as well as the interactions of these variables. In the original AAGIS sample the relationship between TCC and Area varies significantly between the six post-strata, with an overall Rsquared value of approximately 0.46 after the deletion of two outliers. The fixed effects in the prediction model were therefore specified as corresponding to a separate linear fit of TCC in terms of Area in each post-stratum. Random effects (necessary for computation of the EBLUP and the MBDE, but not the M-quantile predictor) were defined as independent regional effects (*i.e.*, a random intercepts specification) on the basis that in the original AAGIS sample the between region variance component explains about 3 per cent of the total residual variability with the two outliers removed. The aim was to estimate the regional averages of TCC.

Tables 5 and 6 show the median relative biases and the median relative RMSEs of different estimators and corresponding estimators of the MSEs of these estimators based on the $K = 1{,}000$ independent stratified samples taken from the Albanian and AAGIS populations respectively. It is noteworthy that in spite of the fact that the linear mixed models fitted to both the Albanian and AAGIS data appear reasonable, the gains from adoption of SAE methods based on them do not lead to substantial improvements in efficiency given the original regional sample sizes for these surveys. On the other hand, the M-quantile estimator, which is not based on a random effects specification, works well both in terms of bias and MSE for the AAGIS population in this case (Table 6, Median $n_i = 53$), while the EBLUP, although the best performer in terms of MSE for the Albanian population (Table 5, Median $n_i = 56$), also records the highest biases (albeit still small, with the largest less than 2%) for both populations given the original area sample sizes. The survey regression estimator performs well, although for both populations there are indirect estimators that perform somewhat better. Design-based simulations based on the Albanian and AAGIS populations were also carried out using smaller area sample sizes than in the original surveys. In particular, the overall sample size was reduced for the Albanian population to $n =$ 291 (with a median district sample size of 9). Similarly, the overall sample size was reduced for the AAGIS population to $n = 233$ (with a median regional sample size of 8). As expected the RMSE of the point estimators increases as the area sample sizes decrease. Overall, the EBLUP improves its RMSE performance relative to all other estimators given these smaller sample sizes. However, since the realism of these reduced sample size designs is somewhat questionable, we do not place too much emphasis on results derived from them, noting only that they are useful for assessing the performance of MSE estimators with realistic data and with very small sample sizes.

Table 5
Performances of estimators of regional means and their MSE estimators – Albanian household population

Weighting Method	**Median $n_i = 56$**		**Median $n_i = 9$**	
Estimator	RB(*m*)	RRMSE(*m*)	RB(*m*)	RRMSE(*m*)
Regression	0.04	6.25	-0.13	16.56
EBLUP, (13)	0.42	5.90	1.62	12.42
MBDE, (18)	0.03	6.14	0.04	16.92
M-quantile, (22)	0.04	6.07	-0.05	16.60
Method/MSE	RB(*M*)	RRMSE(*M*)	RB(*M*)	RRMSE(*M*)
Regression /VReg	17.6	42	11.2	42
EBLUP/PR0	14.6	44	10.5	50
EBLUP/PR1	14.4	43	8.8	48
EBLUP/PR2	14.5	43	9.7	48
EBLUP/Conditional	0.1	24	7.7	99
MBDE/Conditional	-0.8	25	-5.5	64
M-quantile/Conditional	2.9	27	-2.0	75

Table 6
Performances of estimators of regional means and their MSE estimators – AAGIS farm population

Weighting Method	**Median $n_i = 53$**		**Median $n_i = 8$**	
Estimator	RB(*m*)	RRMSE(*m*)	RB(*m*)	RRMSE(*m*)
Regression	0.03	13.36	0.08	29.83
EBLUP, (13)	1.64	13.53	0.92	25.82
MBDE, (18)	-0.73	14.26	-1.02	37.77
M-quantile, (22)	-0.04	11.68	-0.15	32.22
Method/MSE	RB(*M*)	RRMSE(*M*)	RB(*M*)	RRMSE(*M*)
Regression /VReg	74.1	406	54.7	867
EBLUP/PR0	22.4	131	17.7	374
EBLUP/PR1	19.5	137	19.0	367
EBLUP/PR2	21.0	123	31.1	444
EBLUP/Conditional	5.5	132	17.8	255
MBDE/Conditional	-0.5	181	0.9	318
M-quantile/Conditional	-0.7	69	-1.9	212

Focusing on the simulation results obtained using the original regional sample sizes, we see that all three PR-based MSE estimators for the EBLUP display a substantial upward bias in both sets of design-based simulations as well as larger (Albanian population, Table 5) or comparable (AAGIS population, Table 6) instability to the conditional MSE estimators. For the Albanian population all three versions of the conditional MSE estimator are essentially unbiased whereas for the AAGIS population all three versions of the conditional MSE estimator display small or moderate bias.

It is noteworthy that for the Albanian population (Table 5) the relative performances of the PR MSE estimators improve with smaller samples. However, this is because the conditional MSE estimators then become more unstable. For these very small area samples the conditional MSE estimator is less biased than the PR MSE estimator (7.7% *vs*. 10.5%) but is also more unstable (RRMSE of conditional MSE estimator is 99% *vs.* 50% for the PR MSE estimator). This is, however, not the case for the AAGIS population with median $n_i = 8$. In this case, the PR-based MSE estimators perform badly, with the conditional MSE estimators being both less biased and more stable.

The MSE estimator of the regression estimator exhibits moderate or high bias for both populations and all simulation scenarios. For the Albanian population it appears to be competitive to the other MSE estimators in terms of RRMSE but for the AAGIS population it is clearly less stable than the other MSE estimators. Finally, the conditional MSE estimator of the M-quantile estimator performs well with small relative bias and good stability for all simulation scenarios and both populations with the exception of the Albanian population with median $n_i = 9$ where its RRMSE is 75%.

An insight into the reasons for these differences in behaviour can be obtained by examining the area specific RMSE values displayed in Figure 2 for the Albanian population and in Figure 3 for the AAGIS population. Note that in both cases the sample sizes are those from the original surveys. Thus, in Figure 2 we see that all three conditional MSE estimators track the district-specific design-based RMSEs of their respective estimators exceptionally well. In contrast, the PR0 estimator does not seem to be able to capture between district differences in the design-based RMSE of the EBLUP. In Figure 3 we see that the conditional estimator of the MSE of the M-quantile estimator performs extremely well in all regions, with the corresponding estimator of the MSE of the MBDE also performing well in all regions except one (region 6) where it substantially overestimates the design-based RMSE of this predictor. This region is noteworthy because samples that are unbalanced with respect to Area within the region lead to negative weights under the assumed linear mixed model. The picture becomes more complex when one considers the region-specific RMSE estimation performance of the EBLUP in Figure 3. Here we see that the conditional estimator of the MSE of the EBLUP clearly tracks the region-specific design-based RMSE of this predictor better than the PR0 MSE estimator. With the exception of region 6 (where sample balance is a problem), there seems to be little regional variation in the value of the PR0 estimator of the RMSE of the EBLUP, indicating a serious bias problem.

As noted earlier, it is not uncommon to want to produce an estimate for a small area where there is no sample. In such cases, one has to rely completely on the correctness of the model specification. In Table 7 we illustrate the importance of this assumption by contrasting the estimation and MSE estimation performances of the EBLUP for sampled areas with that of the Synthetic EBLUP for areas where no sample data are available. Two situations are shown. The first is a modification of the model-based SIM1-A simulation with a small average sample size and with five zero sample areas. The second is a similar small sample modification of the design-based simulation based on the

AAGIS population, with four zero sample areas. It is clear that when the model underpinning the EBLUP actually holds (*i.e.*, SIM1-A), estimation and MSE estimation (either based on PR0, or on the conditional alternative) works well. The problem is that when there is some doubt about how well this model holds (as in the AAGIS population), then the EBLUP can fail, and our estimator of its MSE can also fail to identify this problem. This is nicely illustrated by the results for the AAGIS population in Table 7 where we see that both the PR0 and conditional MSE estimators for the Synthetic EBLUP completely fail to identify the large positive bias of the Synthetic EBLUP and so end up with a large downward bias.

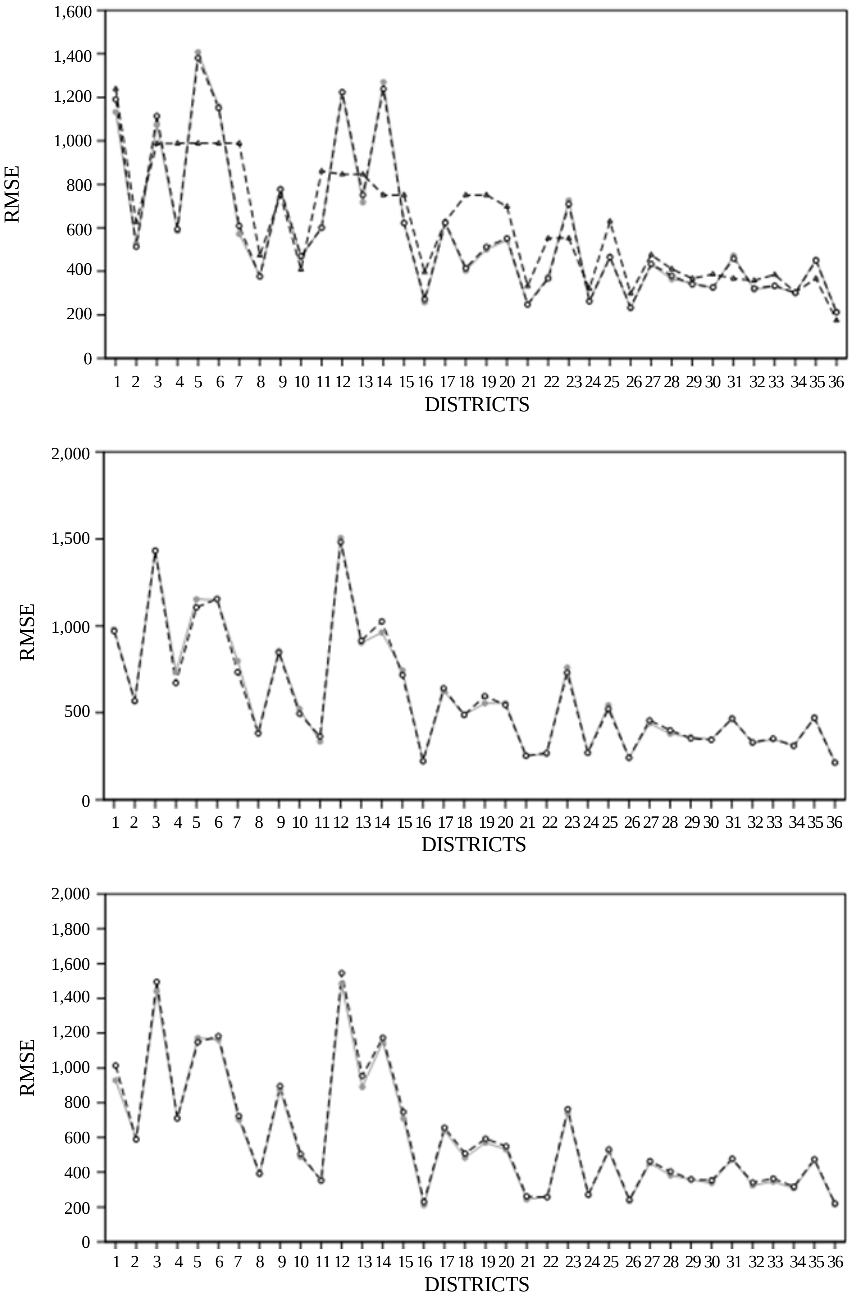


Figure 2 District level values of true design-based RMSE (solid line) and average estimated RMSE (dashed line) obtained in the design-based simulations using the Albanian household population. Districts are ordered in terms of increasing population size. Values for the PR0 estimator are indicated by Δ while those for the conditional estimator are indicated by o. Plots show results for the EBLUP (top), MBDE (centre) and M-quantile (bottom) estimators

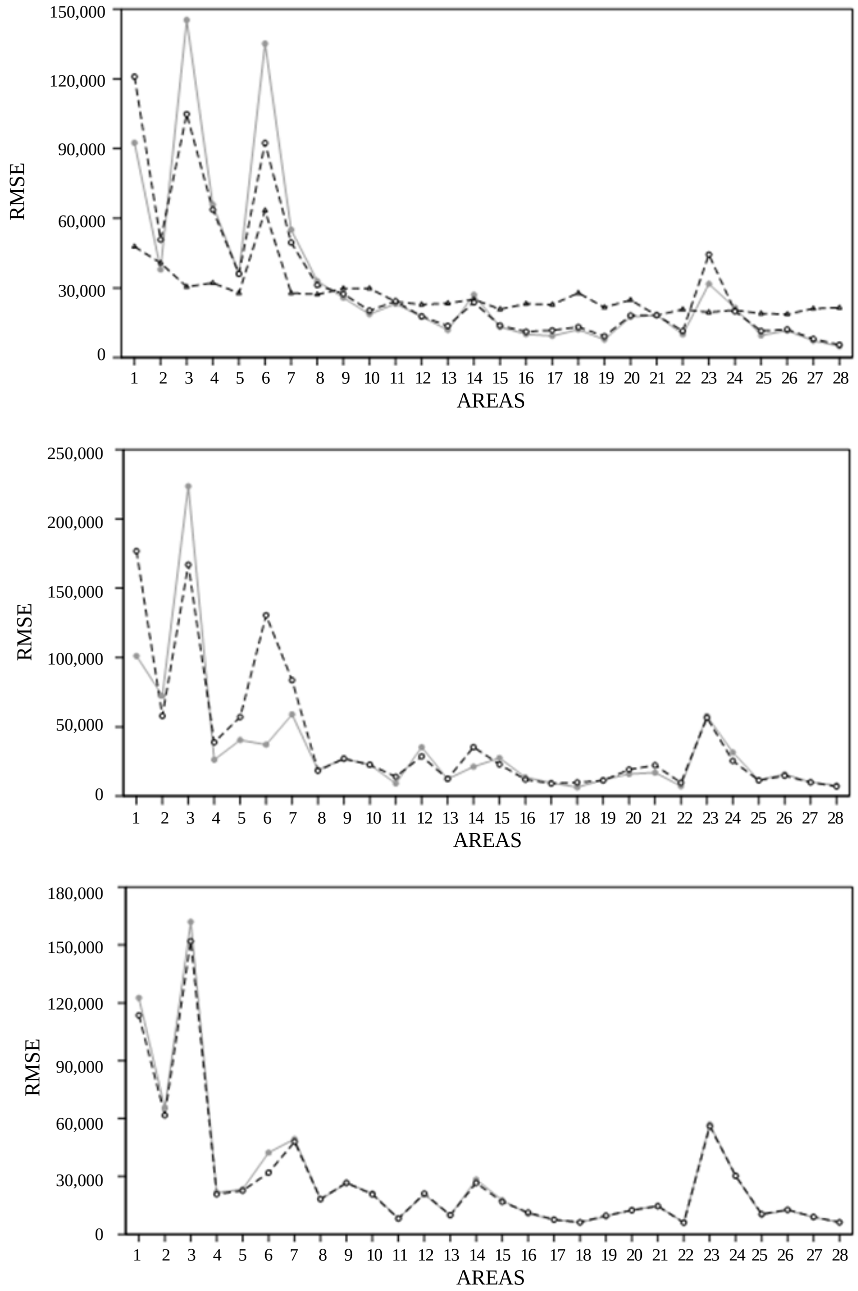


Figure 3 **Regional values of true design-based RMSE (solid line) and average estimated RMSE (dashed line) obtained in the design-based simulations using the AAGIS farm population. Regions are ordered in terms of increasing population size. Values for the PR0 estimator are indicated by Δ while those for the conditional estimator are indicated by o. Plots show results for the EBLUP (top), MBDE (centre) and M-quantile (bottom) estimators**

Table 7
Performance of EBLUP and MSE estimators when there are areas with zero sample

	Weighting Method/ Estimator	SIM1-A, median $n_i = 10$		AAGIS, median $n_i = 9$	
		RB(m)	RRMSE(m)	RB(m)	RRMSE(m)
Areas with $n_i > 0$	(13)/EBLUP	0.00	0.52	2.29	24.94
Areas with $n_i = 0$	(23)/Synthetic EBLUP	-0.05	1.25	87.45	96.46
	MSE Estimator	RB(M)	RRMSE(M)	RB(M)	RRMSE(M)
Areas with $n_i > 0$	(13)/PR0	0.5	11	29.91	760
	(13)/Conditional	0.7	50	23.87	298
Areas with $n_i = 0$	(23)/PR0	-1.8	35	-29.07	601
	(23)/Conditional	-3.6	34	-31.45	101

4. Conclusions and discussion

In this paper we propose a bias-robust and easily implemented method of estimating the conditional MSE of pseudo-linear estimators of small area means (and totals). Our empirical results show that this method of MSE estimation performs reasonably well in terms of bias when used to estimate the model-based MSE and the design-based MSE of the three rather different pseudo-linear estimators considered in this paper. However, this improved bias performance comes at the cost of increased variability. In particular, when area sample sizes are very small, we do not recommend use of our proposed method of MSE estimation for a conditionally biased estimator like the EBLUP.

The EBLUP is widely used in SAE, and in this context the model-dependent MSE estimator PR0 for the EBLUP suggested by Prasad and Rao (1990) is unbiased when its model assumptions are valid (SIM1-A/B and SIM2-A/B in our model-based simulations) but is biased in the presence of outlier area effects (SIM3-A/A* and SIM3-B/B*). It was also the most stable MSE estimator in the model-based simulations. However, its area-averaged construction meant that it did not track the area-specific MSE of the EBLUP in both our design-based simulations, where the correctness of the assumed linear mixed model could only be considered as approximate. This suggests that our proposed conditional MSE estimation method should be considered as an alternative to PR0 in situations where there is some doubt about the correctness of the specification of the small area linear mixed model or where the area sample sizes are not small. Some idea of what constitutes a small sample size can be deduced from the empirical results presented in this paper.

If there is doubt about the validity of the assumed linear mixed model, the user could consider estimation based on a more widely applicable alternative model, *e.g.*, the M-quantile model, or replace the EBLUP by a more outlier-robust alternative (Sinha and Rao 2009). In the former case the approach that we propose in this paper is currently the only analytical approach to MSE estimation, while in the latter case it provides an analytic alternative to more computationally intensive bootstrap methods of MSE estimation. Note however, that for very small area-specific sample sizes the bias-robust MSE estimator proposed in this paper remains unstable.

A future line of research could be to compare the analytic MSE estimation method proposed in this paper with bootstrap-based MSE estimators, *e.g.*, the nonparametric bootstrap MSE estimator of the M-quantile estimator proposed by Tzavidis, Marchetti and Chambers (2010), and the bootstrap MSE estimator for the Robust EBLUP estimator proposed by Sinha and Rao (2009). A key issue in this investigation will be to investigate whether alternative bootstrap MSE estimators are more stable, especially for small area-specific sample sizes.

The extension of the conditional MSE approach to non-linear SAE situations remains to be done. However, since this approach is closely linked to robust population level MSE estimation based on Taylor series linearisation (as well as jackknife estimation of MSE, see Valliant, Dorfman and Royall 2000, section 5.4.2), it should be possible to develop appropriate extensions for corresponding small area non-linear estimation methods. Although the relevant results are not provided here, some evidence for this is that the conditional MSE estimation method described in this paper has already been used to estimate the MSE of the MBDE when it is applied to variables that do not lend themselves to linear mixed modelling, *e.g.*, those with a high proportion of zero values (Chandra and Chambers 2009), and categorical variables (Chandra, Chambers and Salvati 2011). More recently, the approach has also been used to estimate the MSE of geographically weighted M-quantile small area estimators in situations where the small area values are spatially correlated (Salvati, Tzavidis, Pratesi and Chambers 2011). It has also been used by Salvati, Chandra, Ranalli and Chambers (2010) to estimate the MSE of small area estimators based on a nonparametric small area model (Opsomer, Claeskens, Ranalli, Kauermann and Breidt 2008).

As is clear from the development in this paper, our preferred approach to MSE estimation assumes that the MSE of real interest is that defined by the area-specific model (1). This is in contrast to the usual approach to defining MSE in SAE, which adopts an area-averaged MSE concept as the appropriate measure of the accuracy of a small area estimator. As pointed out by Longford (2007), the ultimate aim in SAE is to make inferences about small area characteristics conditional on the realised (but unknown) values of small area effects, *i.e.*, with respect to (1). One can consider this to be a design-based objective (as in Longford 2007), or, as we prefer, a model-based objective that does not quite fit into the usual random effects framework for SAE. In either case we are interested in variability that is with respect to fixed area-specific expected values. This is consistent with the concept of variability that is typically applied in population level inference.

Acknowledgements

The authors would like to acknowledge the valuable comments and suggestions of the Editor, the Associate Editor and two referees. These led to a considerable improvement in the paper.

References

Chambers, R.L., and Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73, 597-604.

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

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

Chandra, H., Chambers, R. and Salvati, N. (2011). Small area estimation of proportions in business surveys. To appear in *Journal of Statistical Computation and Simulation*.

Henderson, C.R. (1953). Estimation of variance and covariance components. *Biometrics*, 9, 226-252.

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

Opsomer, J.D., Claeskens, G., Ranalli, M.G., Kauermann, G. and Breidt, F.J. (2008). Nonparametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society*, Series B, 70, 265-286.

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.

Salvati, N., Chandra, H., Ranalli, M.G. and Chambers, R. (2010). Small area estimation using a nonparametric model based direct estimator. *Computational Statistics and Data Analysis*, 54, 2159-2171.

Salvati, N., Tzavidis, N., Pratesi, M. and Chambers, R. (2011). Small area estimation via M-quantile geographically weighted regression. Forthcoming in *TEST*, DOI 10.1007/s11749-010-0231-1.

Sinha, S.K., and Rao, J.N.K. (2009). Robust small area estimation. *The Canadian Journal of Statistics*, 37(3), 381-399.

Street, J.O, Carroll, R.J. and Ruppert, D. (1988). A note on computing robust regression estimates via iteratively reweighted least squares. *The American Statistician*, 42, 152-154.

Tzavidis, N., Marchetti, S. and Chambers, R. (2010). Robust prediction of small area means and distributions. *Australian and New Zealand Journal of Statistics*, 52, 167-186.

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

Survey Methodology, December 2011
Vol. 37, No. 2, pp. 171-179
Statistics Canada, Catalogue No. 12-001-X

Variance estimation under composite imputation: The methodology behind SEVANI

Jean-François Beaumont and Joël Bissonnette [1]

Abstract

Composite imputation is often used in business surveys. The term "composite" means that more than a single imputation method is used to impute missing values for a variable of interest. The literature on variance estimation in the presence of composite imputation is rather limited. To deal with this problem, we consider an extension of the methodology developed by Särndal (1992). Our extension is quite general and easy to implement provided that linear imputation methods are used to fill in the missing values. This class of imputation methods contains linear regression imputation, donor imputation and auxiliary value imputation, sometimes called cold-deck or substitution imputation. It thus covers the most common methods used by national statistical agencies for the imputation of missing values. Our methodology has been implemented in the System for the Estimation of Variance due to Nonresponse and Imputation (SEVANI) developed at Statistics Canada. Its performance is evaluated in a simulation study.

Key Words: Auxiliary value imputation; Composite imputation; Donor imputation; Imputation model; Linear imputation; Regression imputation; SEVANI.

1. Introduction

Composite imputation is often used in business surveys. The term "composite" means that more than a single imputation method is used to impute missing values for a variable of interest. The choice of a method over another one depends on the availability of auxiliary variables. For instance, ratio imputation could be used to impute a missing value when an auxiliary value is available; otherwise, mean imputation could be an alternative.

The problem of estimating the variance in the presence of a single imputation method has been extensively studied in the literature; *e.g.*, two excellent reviews of this topic are: Lee, Rancourt and Särndal (2001) and Haziza (2009). Although the use of composite imputation occurs frequently in practice, there is little literature on estimating its variance. The literature includes a jackknife variance estimator that was proposed and evaluated empirically in Rancourt, Lee and Särndal (1993). Sitter and Rao (1997) developed further the theory and obtained design-consistent linearization and jackknife variance estimators. In both papers, two imputation methods were considered, with ratio imputation being one of the methods, simple random sampling was used and uniform nonresponse was assumed. Later, Felx and Rancourt (2001) extended the general methodology proposed in Särndal (1992) and Deville and Särndal (1994) to composite imputation using simplifying assumptions. Finally, Shao and Steel (1999) developed an interesting and general reverse approach to variance estimation to deal with composite imputation (see also Kim and Rao 2009). Shao and Steel (1999) claimed that their reverse approach leads to derivations that are less involved than those found in Deville and Särndal (1994). We do not fully agree with this statement. Our results indicate that, in general, our extension to Särndal's approach actually leads to simpler derivations than those obtained with the Shao and Steel approach. The reverse approach may however become quite attractive when the sampling fraction is negligible and a replication variance estimation technique is chosen (see section 7 for greater detail).

We consider the methodology proposed by Särndal (1992) as a starting point. It requires the validity of an imputation model; *i.e.*, a model for the variable being imputed. At first glance, the extension of this methodology to composite imputation seems to be quite tedious, as noted by Shao and Steel (1999), until we notice that most imputation methods used in practice lead to imputed estimators that are linear in the observed values of the variable of interest. This considerably simplifies the derivation of a variance estimator even when there is a single imputation method. For the estimation of the sampling portion of the overall variance, we use a methodology (see Beaumont and Bocci 2009) that is slightly different than the one proposed by Särndal (1992). This allows us to simplify the derivations further. This research has been implemented in version 2 of the System for the Estimation of Variance due to Nonresponse and Imputation (SEVANI), which is developed at Statistics Canada (see Beaumont, Bissonnette and Bocci 2010).

The paper is structured as follows. In section 2, some notation is introduced and composite imputation is explained. Linear imputation is defined in section 3. Our

[1] Jean-François Beaumont, Statistics Canada, Statistical Research and Innovation Division, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6. E-mail: jean-francois.beaumont@statcan.gc.ca; Joël Bissonnette, Statistics Canada, Business Survey Methods Division, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6. E-mail: joel.bissonnette@statcan.gc.ca.

approach to inference and our main assumptions are described in section 4. In section 5, a number of results are stated regarding variance estimation under composite imputation. Section 6 presents the results of a simulation study that assesses the performance of our variance estimator. The reverse approach is briefly discussed in section 7 to highlight the differences with our approach. Finally, a short conclusion is given in section 8.

2. What is composite imputation?

Suppose that we are interested in estimating the population domain total $\theta = \sum_{k \in U} d_k y_k,$ where U is the finite population of size N, y is the variable of interest and d is a domain indicator variable indicating whether population unit k is in the domain of interest $(d_k = 1)$ or not $(d_k = 0)$. A sample s of size n is selected from the finite population U according to a probability sampling design $p(s)$. In the absence of missing values, θ can be estimated by the Horvitz-Thompson estimator $\hat{\theta} = \sum_{k \in s} w_k \, d_k \, y_k,$ where $w_k = 1/\pi_k$ is the design weight and π_k is the selection probability of unit k. Although it is possible to extend our results to calibration estimators, it is not considered in this paper to keep matters simple.

Variable y can be missing for some of the sampled units but we assume that the domain indicator variable d is always observed for those units. The set of sampled units with an observed y-value, called the set of respondents, is denoted by s_r. It is assumed to have been generated according to a nonresponse mechanism $q(s_r \mid s)$. The set of nonrespondents is denoted by $s_m = s - s_r$. It is further split into J mutually exclusive subsets, $s_m^{(j)}, \; j = 1, \ldots, J,$ such that $s_m = \bigcup_{j=1}^{J} s_m^{(j)},$ if composite imputation with $J > 1$ imputation methods is used. All the missing y-values within a given subset $s_m^{(j)}$ are imputed with the same method j. However, different imputation methods are used to impute missing values in different subsets. The resulting imputed estimator can be expressed as

$$\begin{aligned}\hat{\theta}_I &= \sum_{k \in s_r} w_k d_k y_k + \sum_{k \in s_m} w_k d_k y_k^* \\ &= \sum_{k \in s_r} w_k d_k y_k + \sum_{j=1}^{J} \sum_{k \in s_m^{(j)}} w_k d_k y_k^*, \qquad (2.1)\end{aligned}$$

where y_k^* is the imputed y-value for unit k.

Composite imputation is quite frequent in business surveys. It is used because there are missing values in auxiliary variables used for imputation. To fix ideas, let $\mathbf{x}_k$ be the complete vector of auxiliary variables for unit k. Ideally, all the missing y-values would be imputed using a single imputation method based on the complete vector $\mathbf{x}_k$. Unfortunately, there may be missing values in the auxiliary variables so that, for some nonrespondents, we cannot use $\mathbf{x}_k$ to impute their missing y-value; we can only use a subset of $\mathbf{x}_k$. We denote as $\mathbf{x}_k^{\text{obs}}$, the vector of observed auxiliary variables for unit k. This vector does not necessarily contain the same observed variables from one unit to the next. To impute the missing y-value of a given unit k, an imputation method is chosen based on the available auxiliary variables $\mathbf{x}_k^{\text{obs}}$. Since there may be a number of nonresponse patterns in the complete vector of auxiliary variables, the imputation strategy may contain a number of imputation methods.

Example:

The variance estimation issues raised by composite imputation can be better understood by considering the following example. Suppose that the complete vector of auxiliary variables for unit k is $\mathbf{x}_k = (x_{1k}, \, x_{2k}),$ where x_{1k} is strongly related to y_k but subject to missing values while x_{2k} is set to a constant for all sampled units $(x_{2k} = 1, \, k \in s)$. Ideally, x_{1k} is used to impute y_k if it is missing. If x_{1k} is not available, only x_{2k} can be used. Table 1 summarizes the information available for the different subsets of the sample s.

Table 1
Available information when there is one auxiliary variable x_1 and a constant x_2

Subsets		y	x_1	x_2	$\mathbf{x}^{\text{obs}}$
s_r	$s_r^{(1)}$	O	O	O	(x_1, x_2)
	$s_r^{(2)}$	O	M	O	(M, x_2)
s_m	$s_m^{(1)}$	M	O	O	(x_1, x_2)
	$s_m^{(2)}$	M	M	O	(M, x_2)

O: Observed; M: Missing.

The set of nonrespondents s_m is divided into the subsets $s_m^{(1)}$ and $s_m^{(2)}$ depending on the availability of x_1. Similarly, the set of respondents is divided into subsets $s_r^{(1)}$ and $s_r^{(2)}$. In this example, we could use ratio imputation to impute missing y-values in $s_m^{(1)}$ and mean imputation to impute missing y-values in $s_m^{(2)}$. Note that simple linear regression imputation could be used instead of ratio imputation (if it better fits the data). We have chosen ratio imputation in this example for its simplicity and because it is frequently used in business surveys.

Only the respondents in $s_r^{(1)}$ can be used to impute missing y-values in $s_m^{(1)}$ through ratio imputation. The imputed value for a unit k in $s_m^{(1)}$ is $y_k^* = x_{1k} \sum_{l \in s_r^{(1)}} \omega_l^{(1)} y_l \, / \sum_{l \in s_r^{(1)}} \omega_l^{(1)} x_{1l},$ where $\omega_l^{(1)}$ is some weight used for ratio imputation (imputation method 1). Typical choices are: $\omega_l^{(1)} = w_l$ (design-weighted imputation) or $\omega_l^{(1)} = 1$ (unweighted imputation). For mean imputation, the respondents in $s_r^{(2)}$ as well as those in $s_r^{(1)}$ can be used to impute missing y-values in $s_m^{(2)}$. In practice, it is common to use both sets of respondents to improve the stability of the imputed mean. The imputed value for a unit k in $s_m^{(2)}$ is

$$y_k^* = \sum_{l \in s_r} \omega_l^{(2)} y_l \Big/ \sum_{l \in s_r} \omega_l^{(2)},$$

where $\omega_l^{(2)}$ is a weight used for mean imputation (imputation method 2). (Typical choices of $\omega_l^{(2)}$ are the same as those for $\omega_l^{(1)}$; *i.e.*, $\omega_l^{(2)} = w_l$ or $\omega_l^{(2)} = 1$.) This implies that units in $s_r^{(1)}$ can be contributors to both imputation methods. This raises issues for variance estimation of the resulting composite imputation estimator. These issues will be addressed in section 5.

3. What is linear imputation?

The imputation method j is said to be linear if the imputed value y_k^* for a sample unit $k \in s_m^{(j)}$ can be written in the linear form

$$y_k^* = \varphi_{0k}^{(j)} + \sum_{l \in s_r} \varphi_{lk}^{(j)} y_l. \tag{3.1}$$

The quantities $\varphi_{0k}^{(j)}$ and $\varphi_{lk}^{(j)}$, for $l \in s_r$, are obtained without using y-values, but may depend on s and s_r. The linear form (3.1) is satisfied by several of the most common imputation methods in practice such as (weighted or unweighted) linear regression imputation, donor imputation and auxiliary value imputation. A nice review of these methods is found in Haziza (2009). Note that auxiliary value imputation does not use the y-values of respondents; *i.e.*, $y_k^* = \varphi_{0k}^{(j)}$ (see Beaumont, Haziza and Bocci 2011). For donor imputation, the imputed value y_k^* is equal to the y-value of a suitably chosen respondent (donor) so that $\varphi_{0k}^{(j)} = 0$ and $\varphi_{lk}^{(j)} = 0$ for all but one respondent $l \in s_r$. Detailed expressions for $\varphi_{0k}^{(j)}$ and $\varphi_{lk}^{(j)}$ are given in the Methodology Guide of SEVANI (Beaumont, Bissonnette and Bocci 2010), which is available on request from the authors.

Let $\Omega_I^{(j)} = \sum_{k \in s_m^{(j)}} w_k d_k y_k^*$ be the contribution of imputation method j to the estimator $\hat{\theta}_I$. Using (3.1), $\Omega_I^{(j)}$ can be decomposed as follows:

$$\begin{aligned}\Omega_I^{(j)} &= \sum_{k \in s_m^{(j)}} w_k d_k y_k^* \\ &= \sum_{k \in s_m^{(j)}} w_k d_k \varphi_{0k}^{(j)} + \sum_{l \in s_r} y_l \sum_{k \in s_m^{(j)}} w_k d_k \varphi_{lk}^{(j)} \\ &= W_{0d}^{(j)} + \sum_{l \in s_r} W_{dl}^{(j)} y_l ,\end{aligned} \tag{3.2}$$

where $W_{0d}^{(j)} = \sum_{k \in s_m^{(j)}} w_k \, d_k \, \varphi_{0k}^{(j)}$ and $W_{dl}^{(j)} = \sum_{k \in s_m^{(j)}} w_k \, d_k \, \varphi_{lk}^{(j)}$. Using (3.2), the imputed estimator (2.1) can be expressed in the linear form:

$$\begin{aligned}\hat{\theta}_I &= \sum_{k \in s_r} w_k d_k y_k + \sum_{j=1}^{J} \Omega_I^{(j)} \\ &= W_{0d}^{(+)} + \sum_{k \in s_r} (w_k d_k + W_{dk}^{(+)}) \, y_k ,\end{aligned} \tag{3.3}$$

where $W_{0d}^{(+)} = \sum_{j=1}^{J} W_{0d}^{(j)}$ and $W_{dk}^{(+)} = \sum_{j=1}^{J} W_{dk}^{(j)}$.

Continuing with the example introduced at the end of section 2, we observe that, for ratio imputation, $\varphi_{0k}^{(1)} = 0$ and $\varphi_{lk}^{(1)} = \omega_l^{(1)} x_{1k} / \sum_{l \in s_r} \omega_l^{(1)} x_{1l}$, for $l \in s_r$, with $\omega_l^{(1)} = 0$, for $l \in s_r^{(2)}$. For mean imputation, we have $\varphi_{0k}^{(2)} = 0$ and $\varphi_{lk}^{(2)} = \omega_l^{(2)} / \sum_{l \in s_r} \omega_l^{(2)}$, for $l \in s_r$. Consequently, $W_{0d}^{(1)} = 0$, $W_{0d}^{(2)} = 0$,

$$W_{dl}^{(1)} = \omega_l^{(1)} \sum_{k \in s_m^{(1)}} w_k \, d_k \, x_{1k} \Big/ \sum_{k \in s_r} \omega_k^{(1)} x_{1k}$$

and $W_{dl}^{(2)} = \omega_l^{(2)} \sum_{k \in s_m^{(2)}} w_k \, d_k / \sum_{k \in s_r} \omega_k^{(2)}$. This implies that $W_{0d}^{(+)} = 0$ and $W_{dk}^{(+)} = W_{dk}^{(1)} + W_{dk}^{(2)}$.

4. Approach to inference and main assumptions

We consider three sources of variability when evaluating expectations and variances of the imputed estimator: the variability due to the imputation model, the sampling design and the nonresponse mechanism. Note that the use of an imputation model to make inference in the presence of imputation can be found in Rubin (1987), Hidiroglou (1989) and Särndal (1992). In what follows, we will use the subscripts m, p and q to denote the expectations, variances and covariances evaluated with respect to the imputation model, sampling design and nonresponse mechanism respectively.

We consider the following imputation model to describe the relationship between the y-variable and the vector $\mathbf{x}^{\text{obs}}$ of observed auxiliary variables:

$$\begin{aligned}E_m(y_k \mid \mathbf{X}^{\text{obs}}) &= \mu_k \\ V_m(y_k \mid \mathbf{X}^{\text{obs}}) &= \sigma_k^2 \\ \text{cov}_m(y_k, y_l \mid \mathbf{X}^{\text{obs}}) &= 0 ,\end{aligned} \tag{4.1}$$

for $k \neq l$ and $k, l \in U$. The population matrix $\mathbf{X}^{\text{obs}}$ contains the vectors of observed auxiliary variables, $\mathbf{x}_k^{\text{obs}}$, for $k \in U$, and μ_k and σ_k^2 are functions of $\mathbf{x}_k^{\text{obs}}$. Asymptotically m-unbiased and m-consistent estimators of μ_k and σ_k^2 are denoted by $\hat{\mu}_k$ and $\hat{\sigma}_k^2$ respectively. Since we will always condition on $\mathbf{X}^{\text{obs}}$, we exclude this conditioning from the notation to simplify it. For instance, $E_m(y_k \mid \mathbf{X}^{\text{obs}})$ will be written as $E_m(y_k)$.

In model (4.1), we condition on the observed auxiliary variables. Since the nonresponse pattern in the vector $\mathbf{x}$ is not the same for all the nonrespondents, a separate conditional model must be validated and fitted for each nonresponse pattern. In principle, these conditional models should be used to determine the imputation methods chosen.

Note that model (4.1) reduces to the standard conditional model (*e.g.*, Särndal 1992) when the vector $\mathbf{x}$ of auxiliary variables is not subject to missing values.

Remark: The validity of the variance estimation method in section 5 requires μ_k and σ_k^2 to be correctly specified. Although a parametric form for μ_k may often be acceptable, it may be more difficult to determine a suitable parametric form for σ_k^2. To avoid this issue and obtain some robustness against misspecification of the model variance, σ_k^2 can be estimated non parametrically; see the empirical study of Beaumont, Haziza and Bocci (2011) for an illustration of this property under auxiliary value imputation. In the context of donor imputation, Beaumont and Bocci (2009) showed empirically that nonparametric estimation of both μ_k and σ_k^2, via penalized smoothing splines, reduced significantly the vulnerability of our variance estimator to misspecifications of the model mean and variance.

In addition to the imputation model (4.1), we also assume that:

$$F(\mathbf{Y} \mid s, s_r, \mathbf{X}^{\text{obs}}, \mathbf{Z}, \mathbf{D}) = F(\mathbf{Y} \mid \mathbf{X}^{\text{obs}}), \qquad (4.2)$$

where $F(\cdot)$ denotes the distribution function, $\mathbf{Y}$ and $\mathbf{D}$ are N-element vectors containing respectively y_k and d_k as their k^{th} element, and $\mathbf{Z}$ is a N-row matrix of design information, which implicitly or explicitly contains information about the selection probabilities π_k and joint selection probabilities π_{kl}, for $k, l \in U$. This assumption, often implicit in other papers, allows us to treat the response indicators, the domain indicators and the design information as fixed when taking model expectations and variances. A careful choice of the auxiliary variables is necessary to satisfy this assumption. For instance, the design information and the domain indicators should be considered as potential auxiliary variables.

The imputation strategy given in our example started in section 2 could be justified by a model with $\mu_k = \beta_1 x_{1k}$ and $\sigma_k^2 = \sigma_1^2 x_{1k}$, for $k \in s_r^{(1)}$ or $k \in s_m^{(1)}$, and $\mu_k = \beta_2$ and $\sigma_k^2 = \sigma_2^2$, for $k \in s_r^{(2)}$ or $k \in s_m^{(2)}$. The model parameters β_1, β_2, σ_1^2 and σ_2^2 are unknown. Note that if the x_{1k} 's are assumed to be identically distributed random variables with mean μ_x and variance σ_x^2, then $\beta_2 = \beta_1 \mu_x$ and $\sigma_2^2 = \beta_1^2 \sigma_x^2 + \sigma_1^2 \mu_x$. The imputed values $y_k^* = \hat{\mu}_k$, for $k \in s_m$, are obtained by estimating the model parameters β_1 and β_2 from the observed data. For instance, the m-unbiased estimators of β_1 and β_2 could be chosen as

$$\hat{\beta}_1 = \sum\nolimits_{k \in s_r^{(1)}} \omega_k^{(1)} y_k \Big/ \sum\nolimits_{k \in s_r^{(1)}} \omega_k^{(1)} x_{1k}$$

and

$$\hat{\beta}_2 = \sum\nolimits_{k \in s_r^{(2)}} \omega_k^{(2)} y_k \Big/ \sum\nolimits_{k \in s_r^{(2)}} \omega_k^{(2)}$$

respectively. This would lead to $\hat{\mu}_k = \hat{\beta}_1 x_{1k}$, for $k \in s_r^{(1)}$ or $k \in s_m^{(1)}$, and $\hat{\mu}_k = \hat{\beta}_2$, for $k \in s_r^{(2)}$ or $k \in s_m^{(2)}$. As in section 2, one could also consider the potentially more efficient estimator $\hat{\beta}_2^* = \sum_{k \in s_r} \omega_k^{(2)} y_k / \sum_{k \in s_r} \omega_k^{(2)}$ instead of $\hat{\beta}_2$. Unfortunately, $\hat{\beta}_2^*$ is biased under the model since

$$E_m(\hat{\beta}_2^* \mid s, s_r) = \beta_2 + \frac{\sum\limits_{k \in s_r^{(1)}} \omega_k^{(2)} (x_{1k} \beta_1 - \beta_2)}{\sum\limits_{k \in s_r} \omega_k^{(2)}}. \qquad (4.3)$$

As pointed out above, if the x_{1k} 's are assumed to be identically distributed random variables with mean μ_x and variance σ_x^2, $\beta_2 = \beta_1 \mu_x$ and equation (4.3) can be rewritten as

$$E_m(\hat{\beta}_2^* \mid s, s_r) = \beta_2 + \beta_1 \frac{\sum\limits_{k \in s_r^{(1)}} \omega_k^{(2)}}{\sum\limits_{k \in s_r} \omega_k^{(2)}} \frac{\sum\limits_{k \in s_r^{(1)}} \omega_k^{(2)} (x_{1k} - \mu_x)}{\sum\limits_{k \in s_r^{(1)}} \omega_k^{(2)}}. \qquad (4.4)$$

It can be shown under weak conditions that $E_m(\hat{\beta}_2^* \mid s, s_r) = \beta_2 + O_p(1/\sqrt{n})$ so that the model bias of $\hat{\beta}_2^*$ is asymptotically negligible. However, since $\text{var}_m(\hat{\beta}_2^* \mid s, s_r) = O_p(1/n)$, the squared model bias is not necessarily asymptotically negligible compared to the model variance of $\hat{\beta}_2^*$. At least, $\hat{\beta}_2^*$ is m-consistent for β_2. From (4.3) or (4.4), we can see that the model bias of $\hat{\beta}_2^*$ can be controlled by assigning a smaller weight $\omega_k^{(2)}$ to units $k \in s_r^{(1)}$ relative to units $k \in s_r^{(2)}$. For instance, one could consider using $\omega_k^{(2)} = w_k / n^{\alpha}$, for $k \in s_r^{(1)}$ and some $\alpha > 0$, and $\omega_k^{(2)} = w_k$, for $k \in s_r^{(2)}$. In the extreme case where $\omega_k^{(2)} = 0$, for $k \in s_r^{(1)}$, $\hat{\beta}_2^*$ is model-unbiased because it is equal to $\hat{\beta}_2$. Note that the model bias of $\hat{\beta}_2^*$ could be larger than $O_p(1/\sqrt{n})$ if x_{1k}, $k \in s_r^{(1)}$, have a mean different from x_{1k}, $k \in s_r^{(2)}$. In such case, controlling the model bias of $\hat{\beta}_2^*$ might be more important.

In the case of donor imputation, a fourth source of variability needs to be considered when donors are randomly selected among respondents to impute nonrespondents. In this paper, the subscript q will implicitly indicate that moments are evaluated with respect to the joint distribution induced by the nonresponse mechanism and the random donor selection mechanism. As a result, when conditioning on s_r, as in (4.2), it should be kept in mind that conditioning is not only on the set of respondents but also on the set of selected donors.

5. Variance estimation

Särndal (1992) expresses the total error of the imputed estimator as:

$$\hat{\theta}_I - \theta = (\hat{\theta} - \theta) + (\hat{\theta}_I - \hat{\theta}), \qquad (5.1)$$

where the first term on the right-hand side of (5.1) is called the sampling error and the second term is called the nonresponse error. Using the assumptions given in section 4 and $E_p(\hat{\theta} - \theta) = 0,$ the overall bias of the imputed estimator reduces to $E_{mpq}(\hat{\theta}_I - \theta) = E_{pq}B_m,$ where $B_m = E_m(\hat{\theta}_I - \hat{\theta} \mid s, s_r)$ is the (conditional) model bias of the imputed estimator. Using (2.1), the model bias can be expressed as

$$B_m = \sum_{j=1}^{J} \sum_{k \in s_m^{(j)}} w_k d_k E_m(y_k^* - y_k \mid s, s_r). \qquad (5.2)$$

This means that the model bias and the overall bias vanish if the model expectation of the imputation error, $y_k^* - y_k,$ is zero, for $k \in s_m^{(j)}$ and $j = 1, ..., J.$ In principle, an imputation strategy should be chosen so that this condition is satisfied (at least approximately). This is typically assumed in the literature (*e.g.*, Särndal 1992; Shao and Steel 1999).

In the example introduced in section 2, the model bias (5.2) reduces to

$$B_m = \left(\sum_{k \in s_m^{(2)}} w_k d_k \right) E_m(\hat{\beta}_2^* - \beta_2 \mid s, s_r).$$

An expression for $E_m(\hat{\beta}_2^* - \beta_2 \mid s, s_r)$ is given by (4.3) or (4.4). As noted in the paragraph that follows equation (4.4), the model bias, $B_m,$ can be controlled by assigning a smaller weight $\omega_k^{(2)}$ to units $k \in s_r^{(1)}$ relative to units $k \in s_r^{(2)}.$ It is also small if the number of nonrespondents imputed by method 2 is small. Note that our variance (or Mean Squared Error, MSE) estimation approach requires the slightly weaker assumption that $E_q(B_m \mid s)$ is negligible (see section 5.3).

Using (5.1), Särndal (1992) decomposed the overall MSE into three components:

$$E_{mpq}(\hat{\theta}_I - \theta)^2 = E_m \mathrm{var}_p(\hat{\theta}) + E_{pq} E_m\{(\hat{\theta}_I - \hat{\theta})^2 \mid s, s_r\} + 2E_{pq}E_m\{(\hat{\theta}_I - \hat{\theta})(\hat{\theta} - \theta) \mid s, s_r\}. \qquad (5.3)$$

The overall MSE (5.3) becomes approximately equivalent to the overall variance, $\mathrm{var}_{mpq}(\hat{\theta}_I - \theta),$ when the overall bias is negligible. The first, second and third terms on the right-hand side of (5.3) are referred to as the sampling variance, the nonresponse variance and the mixed component respectively. The sum of the last two terms can be called the nonresponse component since these terms would disappear if there were no nonresponse. The nonresponse component is simply the difference between the overall MSE/variance and the sampling variance. In what follows, we develop an estimator for each of these three terms.

5.1 Estimation of the sampling variance

Let $v(y)$ be a p-unbiased estimator of $\mathrm{var}_p(\hat{\theta})$ that would be used under complete response. The typical Horvitz-Thompson estimator is

$$v(y) = \sum_{k \in s} \sum_{l \in s} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} (w_k d_k y_k)(w_l d_l y_l), \qquad (5.4)$$

where π_{kl} is the joint selection probability of units k and l. In the presence of nonresponse, $\hat{V}_{\mathrm{ORD}} = v(y_\bullet)$ is the naïve sampling variance estimator that treats the imputed values as true values, where $y_\bullet$ is the imputed y-variable; *i.e.*, $y_{\bullet k} = y_k,$ for $k \in s_r,$ and $y_{\bullet k} = y_k^*,$ for $k \in s_m.$

Särndal (1992) proposed the following *mpq*-unbiased estimator of the sampling variance $V_{\mathrm{SAM}} = E_m \mathrm{var}_p(\hat{\theta})$:

$$\hat{V}_{\mathrm{SAM}} = \hat{V}_{\mathrm{ORD}} + \hat{V}_{\mathrm{DIF}},$$

where $\hat{V}_{\mathrm{DIF}}$ is an m-unbiased estimator of $V_{\mathrm{DIF}} = E_m(v(y) - \hat{V}_{\mathrm{ORD}} \mid s, s_r).$ Unfortunately, the expression for $\hat{V}_{\mathrm{DIF}}$ is usually tedious to derive, and it is even more so when composite imputation is used.

Beaumont and Bocci (2009) simplified Särndal's derivations by conditioning on $\mathbf{Y}_r,$ the vector containing the responding y-values. More explicitly, let $V_{\mathrm{DIF}}^{\mathrm{C}} = E_m(v(y) - \hat{V}_{\mathrm{ORD}} \mid s, s_r, \mathbf{Y}_r)$ and $\hat{V}_{\mathrm{DIF}}^{\mathrm{C}}$ be an m-unbiased estimator of $V_{\mathrm{DIF}}^{\mathrm{C}}$; *i.e.*, $E_m(\hat{V}_{\mathrm{DIF}}^{\mathrm{C}} \mid s, s_r, \mathbf{Y}_r) = V_{\mathrm{DIF}}^{\mathrm{C}}.$ Our *mpq*-unbiased sampling variance estimator is $\hat{V}_{\mathrm{SAM}}^{\mathrm{C}} = \hat{V}_{\mathrm{ORD}} + \hat{V}_{\mathrm{DIF}}^{\mathrm{C}}.$ Since $\hat{V}_{\mathrm{ORD}}$ is a constant when conditioning on $s,$ s_r and $\mathbf{Y}_r,$ $\hat{V}_{\mathrm{SAM}}^{\mathrm{C}}$ can simply be obtained by estimating $E_m(v(y) \mid s, s_r, \mathbf{Y}_r).$ If (5.4) is used,

$$E_m(v(y) \mid s, s_r, \mathbf{Y}_r) = v(y_\bullet^\mu) + \sum_{k \in s_m} (1 - \pi_k) w_k^2 d_k \sigma_k^2, \qquad (5.5)$$

where $y_{\bullet k}^\mu = y_k,$ for $k \in s_r,$ and $y_{\bullet k}^\mu = \mu_k,$ for $k \in s_m.$ An estimator $\hat{V}_{\mathrm{SAM}}^{\mathrm{C}}$ of (5.5) is obtained by replacing the unknown mean μ_k and unknown variance σ_k^2 in (5.5) by m-unbiased (or at least m-consistent) estimators $\hat{\mu}_k$ and $\hat{\sigma}_k^2.$ This estimator is easy to compute provided a software package that treats the complete response case is available to obtain the first term on the right-hand side of (5.5). The general formula (5.5) can be used for every imputation strategy. The only difference between different imputation strategies lies in the choice of the imputation model and the estimators $\hat{\mu}_k$ and $\hat{\sigma}_k^2.$

5.2 Estimation of the nonresponse variance

An *mpq*-unbiased estimator of the nonresponse variance $V_{\mathrm{NR}} = E_{pq} E_m\{(\hat{\theta}_I - \hat{\theta})^2 \mid s, s_r\}$ is obtained by finding an m-unbiased estimator of

$$E_m\{(\hat{\theta}_I - \hat{\theta})^2 \mid s, s_r\} = \text{var}_m\{(\hat{\theta}_I - \hat{\theta}) \mid s, s_r\} + B_m^2. \quad (5.6)$$

Using $\hat{\theta}_I$ defined in the first equation of (3.3), the nonresponse error with composite imputation can be decomposed into J components:

$$\hat{\theta}_I - \hat{\theta} = \sum_{j=1}^{J} (\Omega_I^{(j)} - \Omega^{(j)}),$$

where $\Omega^{(j)} = \sum_{k \in s_m^{(j)}} w_k d_k y_k$. Each of these J components, $\Omega_I^{(j)} - \Omega^{(j)}$, is associated with a different imputation method. Since y_k^* only involves observed y-values, $\Omega_I^{(j)} = \sum_{k \in s_m^{(j)}} w_k d_k y_k^*$ only involves observed y-values as well and thus $\Omega_I^{(j)}$ and $\Omega^{(j)}$ are independent under the model. Therefore, the model variance of the nonresponse error can be written as

$$\text{var}_m\{(\hat{\theta}_I - \hat{\theta}) \mid s, s_r\} = \sum_{i=1}^{J}\sum_{j=1}^{J} \text{cov}_m(\Omega_I^{(i)}, \Omega_I^{(j)} \mid s, s_r) + \sum_{j=1}^{J} \text{var}_m(\Omega^{(j)} \mid s, s_r). \quad (5.7)$$

Note that the covariances $\text{cov}_m(\Omega_I^{(i)}, \Omega_I^{(j)} \mid s, s_r)$, for $i \neq j$, are not necessarily negligible because some observed y-values can be used for more than one imputation method.

The derivations of the model variance (5.7) could be quite involved when several imputation methods are used because of the non-negligible covariances. The algebra can be greatly simplified for linear imputation methods. By using the second equation given in (3.3), the nonresponse error can be expressed as

$$\hat{\theta}_I - \hat{\theta} = W_{0d}^{(+)} + \sum_{k \in s_r} W_{dk}^{(+)} y_k - \sum_{k \in s_m} w_k d_k y_k. \quad (5.8)$$

Since the nonresponse error is linear in the y-values, its model variance is given by

$$\text{var}_m\{(\hat{\theta}_I - \hat{\theta}) \mid s, s_r\} = \sum_{k \in s_r} (W_{dk}^{(+)})^2 \sigma_k^2 + \sum_{k \in s_m} w_k^2\, d_k\, \sigma_k^2. \quad (5.9)$$

If the model bias B_m is negligible, an *mpq*-unbiased estimator $\hat{V}_{\text{NR}}$ of the nonresponse variance V_{NR} is obtained by replacing σ_k^2 in (5.9) by an *m*-unbiased (and *m*-consistent) estimator $\hat{\sigma}_k^2$. If the model bias is not negligible, it can be estimated by an *m*-consistent estimator $\hat{B}_m$ and, using equation (5.6), the nonresponse variance estimator $\hat{V}_{\text{NR}}$ can be replaced by $\hat{V}_{\text{NR}} + \hat{B}_m^2$. Note that $\hat{B}_m^2$ is *m*-consistent for B_m^2 provided that $\hat{B}_m$ is *m*-consistent for B_m. The estimator $\hat{B}_m$ can be found by using (5.8) and writing the model bias as

$$\begin{aligned} B_m &= E_m(\hat{\theta}_I - \hat{\theta} \mid s, s_r) \\ &= W_{0d}^{(+)} + \sum_{k \in s_r} W_{dk}^{(+)} \mu_k - \sum_{k \in s_m} w_k\, d_k\, \mu_k. \end{aligned} \quad (5.10)$$

The estimator $\hat{B}_m$ is obtained by replacing μ_k in (5.10) by an *m*-consistent estimator $\hat{\mu}_k$.

5.3 Estimation of the mixed component

An *mpq*-unbiased estimator of the mixed component

$$V_{\text{MIX}} = 2E_{pq}E_m\{(\hat{\theta}_I - \hat{\theta})(\hat{\theta} - \theta) \mid s, s_r\}$$

is obtained by finding an *m*-unbiased estimator of

$$\begin{aligned} 2E_m\{(\hat{\theta}_I - \hat{\theta})(\hat{\theta} - \theta) \mid s, s_r\} = {} & \\ & 2\,\text{cov}_m\{(\hat{\theta}_I - \hat{\theta}), (\hat{\theta} - \theta) \mid s, s_r\} \\ & + 2B_m E_m\{(\hat{\theta} - \theta) \mid s, s_r\}. \end{aligned} \quad (5.11)$$

Since both the nonresponse error and the sampling error are linear in the y-values, using (5.8) we obtain:

$$2\,\text{cov}_m\{(\hat{\theta}_I - \hat{\theta})(\hat{\theta} - \theta) \mid s, s_r\} = 2\sum_{k \in s_r} W_{dk}^{(+)} (w_k - 1)\, d_k \sigma_k^2 - 2\sum_{k \in s_m} w_k (w_k - 1)\, d_k \sigma_k^2. \quad (5.12)$$

If the model bias B_m is negligible, an *mpq*-unbiased estimator $\hat{V}_{\text{MIX}}$ of the mixed component V_{MIX} is obtained by replacing σ_k^2 in (5.12) by an *m*-unbiased (and *m*-consistent) estimator $\hat{\sigma}_k^2$. Note that the mixed component is not necessarily negligible (Brick, Kalton and Kim 2004) and, moreover, it has been found to often be negative in practice.

If the model bias B_m is not negligible, it may not be possible to easily estimate the second component on the right-hand side of (5.11). The reason is that $E_m\{(\hat{\theta} - \theta) \mid s, s_r\}$ involves knowing $\mathbf{x}_k^{\text{obs}}$ as well as the domain indicator variable d for the nonsampled portion of the population; this information may not be available. This problem can be bypassed by changing the inferential framework. The full multivariate distribution between y, $\mathbf{x}$ and d can be modeled instead of conditioning on d and $\mathbf{x}^{\text{obs}}$. We did not implement this idea in SEVANI because it leads to a more complex modeling task and makes it difficult to obtain a general variance expression that is easy to implement. Ignoring the second component on the right-hand side of (5.11) should not be of great concern in practice when the model bias is not too large. In section 5.4, we provide a diagnostic that can be helpful for determining whether the model bias is important or not.

The mixed component can also be written as

$$\begin{aligned} V_{\text{MIX}} &= 2E_{pq}E_m\{(\hat{\theta}_I - \hat{\theta})(\hat{\theta} - \theta) \mid s,\, s_r\} \\ &= 2E_{pq}[\text{cov}_m\{(\hat{\theta}_I - \hat{\theta}),\, (\hat{\theta} - \theta) \mid s, s_r\}] \\ &\quad + 2E_p[E_q(B_m \mid s)\, E_m\{(\hat{\theta} - \theta) \mid s\}]. \end{aligned}$$

Expression (5.12) can therefore be used to obtain an estimator of V_{MIX} provided that $E_q(B_m \mid s)$ is negligible. This is a weaker assumption than requiring B_m to be negligible since this assumption is satisfied when either B_m or $E_q(\hat{\theta}_I - \hat{\theta} \mid s)$ is negligible. For instance, in our earlier example, B_m may not be negligible but, if $d_k = 1$ and $\omega_k^{(1)} = \omega_k^{(2)} = w_k$, $E_q(\hat{\theta}_I - \hat{\theta} \mid s) \approx 0$ under uniform nonresponse (see Sitter and Rao 1997).

5.4 Estimation of the overall MSE/variance

The overall MSE, or overall variance if the overall bias is negligible,

$$V_{\text{TOT}} = E_{mpq}(\hat{\theta}_I - \theta)^2 = V_{\text{SAM}} + V_{\text{NR}} + V_{\text{MIX}}$$

can be estimated by $\hat{V}_{\text{TOT}} = \hat{V}_{\text{SAM}}^{\text{C}} + \hat{V}_{\text{NR}} + \hat{V}_{\text{MIX}}$ if the model bias, B_m, is negligible. The nonresponse component estimator is $\hat{V}_{\text{NR}} + \hat{V}_{\text{MIX}}$. From a user's perspective, the estimator $\hat{V}_{\text{TOT}}$ is of greater interest than its individual components. A user may nevertheless be interested in the estimator of the sampling variance, $\hat{V}_{\text{SAM}}^{\text{C}}$, or the ratio $\hat{V}_{\text{SAM}}^{\text{C}} / \hat{V}_{\text{TOT}}$. The latter estimates the contribution of the sampling variance to the overall variance.

As pointed out in section 5.2, if the model bias is not negligible, the nonresponse variance can be estimated by $\hat{V}_{\text{NR}} + \hat{B}_m^2$ instead of $\hat{V}_{\text{NR}}$. This leads to the overall MSE estimator $\hat{V}_{\text{TOT, ADJ}} = \hat{V}_{\text{SAM}}^{\text{C}} + (\hat{V}_{\text{NR}} + \hat{B}_m^2) + \hat{V}_{\text{MIX}}$.

A statistic that can be useful as a diagnostic to determine the magnitude of the model bias is either $|\hat{B}_m| / \sqrt{\hat{V}_{\text{TOT}}}$ or $|\hat{B}_m| / \sqrt{\hat{V}_{\text{TOT, ADJ}}}$. A large value of any of these two statistics may be an indication that the model bias is not negligible and that the composite imputation procedure should be questioned. The advantage of $|\hat{B}_m| / \sqrt{\hat{V}_{\text{TOT, ADJ}}}$ over $|\hat{B}_m| / \sqrt{\hat{V}_{\text{TOT}}}$ is that it is bounded; *i.e.*,

$$0 \le |\hat{B}_m| / \sqrt{\hat{V}_{\text{TOT, ADJ}}} \le 1.$$

5.5 Random regression imputation

A random regression residual e_k is sometimes added to the regression imputed value y_k^* to preserve the natural variability of the *y*-variable. We suggest that the random residuals e_k be generated independently with $E_*(e_k \mid s, s_r) = 0$ and $\text{var}_*(e_k \mid s, s_r) = \hat{\sigma}_k^2$, where the subscript * indicates that the expectation and variance are taken with respect to the random imputation mechanism. This leads to the imputed value $y_k^{*R} = y_k^* + r_k e_k$, with $r_k = 1$ if unit *k* has been imputed with a random residual added and $r_k = 0$ otherwise. The imputed estimator (2.1) with y_k^* replaced by y_k^{*R} is denoted by $\hat{\theta}_I^* = \hat{\theta}_I + \sum_{k \in s_m} w_k d_k r_k e_k$. Since $E_*(e_k \mid s, s_r) = 0$, adding a random residual does not introduce any bias in the imputed estimator. The overall MSE of $\hat{\theta}_I^*$ can be expressed as

$$E_{mpq*}(\hat{\theta}_I^* - \theta)^2 = E_{mpq}(\hat{\theta}_I - \theta)^2 + E_{mpq}\text{var}_*(\hat{\theta}_I^* \mid s,\, s_r). \quad (5.13)$$

The first term on the right-hand side of (5.13) is estimated as in section 5.4. The second term is estimated by

$$\text{var}_*(\hat{\theta}_I^* \mid s,\, s_r) = \sum_{k \in s_m} w_k^2\, d_k\, r_k\, \hat{\sigma}_k^2. \quad (5.14)$$

6. Simulation study

We conducted a Monte-Carlo simulation study to assess the methodology described in section 5. A bivariate population of $N = 400$ units was generated that contains an auxiliary variable *x* and a variable of interest *y*. For each population unit, the auxiliary variable was generated according to a gamma distribution with mean 48 and variance 768. The variable of interest *y* was generated conditionally on *x* from a gamma distribution with mean $1.5x$ and variance $16x$. Half of the population was randomly assigned a missing value to *x*. As no domain of interest was generated, θ is the overall population total of variable *y*.

Ten thousand samples were selected from this population using simple random sampling without replacement. We considered two sample sizes: $n = 100$ and $n = 250$. For each sample, nonresponse to variable *y* was generated independently from one unit to another with a nonresponse probability of 0.3. We used the same imputation strategy as in the example in section 2 with $\omega_l^{(1)} = 1$, for $l \in s_r^{(1)}$, and $\omega_l^{(2)} = 1$, for $l \in s_r$. Nonrespondents to variable *y* with an observed *x*-value were imputed by ratio imputation while those with a missing *x*-value were imputed by mean imputation.

The population *y*-values were kept fixed throughout the replications of the simulation experiment; each replication consisted of selecting a sample and then generating nonresponse to variable *y*. If we had strictly followed the theoretical development in section 5, we would have generated new *y*-values at each replication according to the imputation model. However, it is more common in the literature to fix the population *y*-values when conducting a simulation experiment. For instance, our simulation set-up is essentially the same as the one discussed in Rancourt, Lee and Särndal (1993), who also considered composite imputation.

We computed the Monte-Carlo sampling variance and overall MSE as $V_{\text{SAM}}^{\text{MC}} = \sum_{r=1}^{R} (\hat{\theta}_r - \theta)^2 / R$ and $V_{\text{TOT}}^{\text{MC}} = \sum_{r=1}^{R} (\hat{\theta}_{I,\,r} - \theta)^2 / R$ respectively, where the subscript r indicates that estimates are computed using the r^{th} replicate and $R = 10{,}000$. The Monte-Carlo relative bias of any estimator of V_{SAM}, say v_{SAM}, is computed as $\text{RB}(V_{\text{SAM}}) = \sum_{r=1}^{R} (v_{\text{SAM},r} - V_{\text{SAM}}^{\text{MC}}) \,/\, (V_{\text{SAM}}^{\text{MC}} R)$. Similarly, we computed the Monte-Carlo relative bias of an estimator of V_{TOT}, denoted as $\text{RB}(V_{\text{TOT}})$, and the Monte-Carlo relative bias of an estimator of $V_{\text{SAM}} / V_{\text{TOT}}$, denoted as $\text{RB}(V_{\text{SAM}} / V_{\text{TOT}})$. Finally, we computed the Monte-Carlo coverage rates of confidence intervals for θ with a 95% confidence level assuming that $\hat{\theta}_I$ is normally distributed.

The results of our simulation study are given in table 2. In the columns labeled SEVANI, the sampling variance, V_{SAM}, and the overall MSE, V_{TOT}, are estimated for each sample by $\hat{V}_{\text{SAM}}^{\text{C}}$ and $\hat{V}_{\text{TOT, ADJ}}$ respectively (see section 5.4). We have also obtained results by replacing $\hat{V}_{\text{TOT, ADJ}}$ by $\hat{V}_{\text{TOT}}$. We do not report these additional results in table 2 as they were quite close to those obtained with $\hat{V}_{\text{TOT, ADJ}}$. This suggests that the model bias B_m is not important in this case. In the columns labeled Naïve, both the sampling variance and the overall MSE are estimated by $\hat{V}_{\text{ORD}}$ (see section 5.1).

Table 2
Results of the simulation study

	$n = 100$		$n = 250$	
	SEVANI	**Naïve**	**SEVANI**	**Naïve**
$\text{RB}(V_{\text{SAM}})$	2.82%	-17.59%	3.02%	-17.68%
$\text{RB}(V_{\text{SAM}}/V_{\text{TOT}})$	8.30%	-	5.84%	-
$\text{RB}(V_{\text{TOT}})$	-5.07%	-40.68%	-2.66%	-52.89%
Coverage Rate	93.38%	86.20%	94.42%	81.80%

These results show that the methodology described in section 5 and implemented in SEVANI is better than the naïve variance estimator for the estimation of the components of variance and the construction of confidence intervals. The use of SEVANI leads to small Monte-Carlo relative biases and coverage rates close to the targeted nominal rate (95%). Our methodology is also useful for users who would like to estimate the contribution of the sampling variance to the overall MSE; *i.e.*, $V_{\text{SAM}} / V_{\text{TOT}}$. Note that $V_{\text{SAM}}^{\text{MC}} / V_{\text{TOT}}^{\text{MC}}$ is 71.98% for $n = 100$ and 57.23% for $n = 250$. Since $V_{\text{SAM}}^{\text{MC}} / V_{\text{TOT}}^{\text{MC}}$ is not close to 100% even for $n = 100$, the effects of nonresponse and imputation cannot be systematically ignored when estimating the overall MSE.

7. The reverse approach

Shao and Steel (1999) proposed a reverse approach to variance estimation developed to deal with composite imputation. They assumed that the overall bias is negligible and suggested the following decomposition of the overall variance:

$$E_{mpq}(\hat{\theta}_I - \theta)^2 = E_{mq}\text{var}_p(\hat{\theta}_I \mid U_r) + E_{mq}\{E_p(\hat{\theta}_I \mid U_r) - \theta\}^2, \qquad (7.1)$$

where U_r is a conceptual population of respondents. The inner expectation and variance in the right side of (7.1) are taken with respect to the sampling design. Unfortunately, the imputed estimator $\hat{\theta}_I$ is generally not linear with respect to the sampling design even though it is linear with respect to the observed y-values. Therefore, the imputed estimator $\hat{\theta}_I$ is typically linearized (*e.g.*, Shao and Steel 1999; Kim and Rao 2009). More explicitly, the quantities $\varphi_{0k}^{(j)}$ and $\varphi_{lk}^{(j)}$ often depend on the sample in a nonlinear way; *e.g.*, this is true with linear regression imputation (see the example at the end of section 3) and donor imputation. It is not always straightforward to account for the sampling variability of $\varphi_{0k}^{(j)}$ and $\varphi_{lk}^{(j)}$ when using (7.1). For example, there is no literature on the use of the reverse approach to estimate the variance under nearest-neighbour imputation. Moreover, since each composite imputation strategy yields its own linearized imputed estimator, it is not an easy task to implement this methodology in a generalized software package.

Using our approach, the inner expectation in the expressions for the nonresponse variance,

$$V_{\text{NR}} = E_{pq}E_m\{(\hat{\theta}_I - \hat{\theta})^2 \,|\, s, s_r\},$$

and the mixed component,

$$V_{\text{MIX}} = 2E_{pq}E_m\{(\hat{\theta}_I - \hat{\theta})(\hat{\theta} - \theta) \mid s, s_r\},$$

are taken with respect to the imputation model (conditionally on s and s_r). The imputed estimator is linear and the derivations are straightforward because the quantities $\varphi_{0k}^{(j)}$ and $\varphi_{lk}^{(j)}$ are constructed without using the y-values. The estimation of the sampling variance, $V_{\text{SAM}} = E_m \text{var}_p(\hat{\theta})$, does not involve these two quantities (see equation 5.5); thus, their possible non-linearity with respect to the sampling design does not cause any difficulty. This implies that nearest-neighbour imputation can be easily handled with our approach (see Beaumont and Bocci 2009).

It is for all the above reasons that we believe that the reverse approach might be more cumbersome to implement in a generalized software package than our approach. This does not mean that the reverse approach is not useful. Indeed, both approaches lead to identical variance estimators when a census is conducted. Beaumont, Haziza and Bocci (2011) showed that they also lead to identical variance estimators under auxiliary value imputation (because $\varphi_{0k}^{(j)}$ and $\varphi_{lk}^{(j)}$ do not depend on s and s_r). Both approaches depend on the correct specification of the imputation model and no approach is expected to systematically outperform the other.

The reverse approach may have a practical advantage over our approach when the sampling fraction is negligible. In such case, Shao and Steel (1999) showed that the second component on the right side of (7.1) can be neglected. The first component is estimated by finding a design-based estimator of $\text{var}_p(\hat{\theta}_I \mid U_r)$. If a replication variance estimation technique (*e.g.*, the jackknife or the bootstrap) is chosen for the estimation of $\text{var}_p(\hat{\theta}_I \mid U_r)$, the whole approach becomes quite attractive and practical. Also, it does not depend on the validity of the imputation model; in particular, the correct specification of the model variance σ_k^2. The jackknife variance estimators of Rancourt, Lee and Särndal (1993) and Sitter and Rao (1997) can be justified by this approach.

8. Conclusion

Our methodology for composite imputation has been implemented in version 2 of SEVANI because of its ease of implementation and generality. It works for most imputation methods used in practice, as most imputation methods are linear. The variance computations are the same for every composite imputation strategy once the quantities $W_{0d}^{(+)}$, $W_{dk}^{(+)}$, $\hat{\mu}_k$ and $\hat{\sigma}_k^2$ have been computed. This eases the development of a generalized system.

Although we have focused on the estimation of a domain total using the Horvitz-Thompson estimator, SEVANI can also deal with domain means and calibration estimators. Parametric and nonparametric methods of estimating μ_k and σ_k^2 are also available. Greater detail can be found in the Methodology Guide of SEVANI (Beaumont, Bissonnette and Bocci 2010) available upon request from the authors.

Acknowledgements

We would like to thank the reviewers for their comments. We would also like to thank Mike Hidiroglou, Eric Rancourt and Cynthia Bocci from Statistics Canada for their suggestions and discussions on the topic. All these comments contributed to improve the paper.

References

Beaumont, J.-F., Bissonnette, J. and Bocci, C. (2010). SEVANI, version 2.3, Methodology Guide. Internal report, Methodology Branch, Statistics Canada.

Beaumont, J.-F., and Bocci, C. (2009). Variance estimation when donor imputation is used to fill in missing values. *Canadian Journal of Statistics*, 37, 400-416.

Beaumont, J.-F., Haziza, D. and Bocci, C. (2011). On variance estimation under auxiliary value imputation in sample surveys. *Statistica Sinica*, 21, 515-537.

Brick, J.M., Kalton, G. and Kim, J.K. (2004). Variance estimation with hot deck imputation using a model. *Survey Methodology*, 30, 57-66.

Deville, J.-C., and Särndal, C.-E. (1994). Variance estimation for the regression imputed Horvitz-Thompson estimator. *Journal of Official Statistics*, 10, 381-394.

Felx, P., and Rancourt, E. (2001). Applications of Variance due to Imputation in the Survey of Employment, Payrolls and Hours. Methodology Branch Working Paper, Statistics Canada, BSMD-2001-009E.

Haziza, D. (2009). Imputation and inference in the presence of missing data. In *Handbook of Statistics, Sample Surveys: Theory, Methods and Inference*, (Eds., D. Pfeffermann and C.R. Rao). Amsterdam: Elsevier BV, 29A, 215-246.

Hidiroglou, M.A. (1989). Unpublished handwritten notes kindly shared with us by the author.

Kim, J.-K., and Rao, J.N.K. (2009). Unified approach to linearization variance estimation from survey data after imputation for item nonresponse. *Biometrika*, 96, 917-932.

Lee, H., Rancourt, E. and Särndal, C.-E. (2001). Variance estimation from survey data under single imputation. In *Survey Nonresponse*, (Eds., R.M. Groves, D.A. Dillman, J.L. Eltinge and R.J.A. Little). New-York: John Wiley & Sons, Inc., 315-328.

Rancourt, E., Lee, H. and Särndal, C.-E. (1993). Variance estimation under more than one imputation method. In *Proceedings of the International Conference on Establishments Surveys*, June 1993, Buffalo, American Statistical Association, 374-379.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New-York: John Wiley & Sons, Inc.

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

Shao, J., and Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94, 254-265.

Sitter, R.R., and Rao, J.N.K. (1997). Imputation for missing values and corresponding variance estimation. *Canadian Journal of Statistics*, 25, 61-73.

Survey Methodology, December 2011
Vol. 37, No. 2, pp. 181-182
Statistics Canada, Catalogue No. 12-001-X

Alternative demographic sample designs being explored at the U.S. Census Bureau

Patrick E. Flanagan and Ruth Ann Killion [1]

1. Introduction

The United States (U.S.) Census Bureau Demographic Survey Sample Redesign Program, among other things, is responsible for research into improving the designs of U.S. demographic surveys, particularly focused on the design of survey sampling. Historically, the research into improving sample design has been restricted to the "mainstream" methods like basic stratification, multi-stage designs, systematic sampling, probability-proportional-to-size sampling, clustering, and simple random sampling. Over the past thirty years or more, we have increasingly faced reduced response rates and higher costs coupled with an increasing demand for more data on all types of populations. More recently, dramatic increases in computing power and availability of auxiliary data from administrative records have indicated that we may have more options than we did when we established our current methodology. Thus, we began an initiative to explore alternative sampling methods.

2. History of innovation in demographic survey sampling at the U.S. Census Bureau

The U.S. Census Bureau was created by the Permanent Census Act of 1902. Up until the late 1930s, the U.S. Census Bureau's demographic work was mostly focused on the logistics of running each decennial census and a myriad of special censuses. After the 1930 decennial census, the Census Bureau began research into sampling using the census data (Stephan 1948).

Then, in 1937, the Census Bureau took its first major step into sample survey sampling with the 1937 Enumerative Check Census of Unemployment, which used a cluster sample of counties in support of a register census of the unemployed (Dedrick 1938). About the same time, the Census Bureau brought in sampling experts (*e.g.*, W. Edwards Deming and Federick Stephan) in its decennial census expansion to assist in designing a sample survey in conjunction with the 1940 Decennial Census using a five percent systematic sample (Stephan, Deming and Hansen 1940). In 1942, the Sample Survey of Unemployment was moved from the Works Progress Administration to the Census Bureau. This survey was already a three-stage sample with county primary sampling units (PSUs), systematic sampling of blocks, and sampling listed housing units in stage three (Frankel and Stock 1942). After its transfer to the Census Bureau (and a name change to the Monthly Report on the Labor Force (MRLF)), it was extensively redesigned in 1943, dramatically improving its efficiency using larger primary sampling units (PSUs) and probability proportionate to size for selection (Duncan and Shelton 1978). Later the survey was changed to improve month-to-month and year-to-year comparisons using a more complex overlapping sample approach in which a given household remains in sample for four months, is out of the survey for eight months and then is back into the sample for four months. Its name was also changed in 1947 to the Current Population Survey (CPS). Still, the basic sampling concept remained multi-stage sample design with county or county group PSUs. It remains that way to present though there are vast differences in the within-PSU sampling methods (U.S. Bureau of Labor Statistics and U.S. Census Bureau 2006). Over the last 60 years, the U.S. Census Bureau has designed many additional demographic surveys. Some of those surveys use the same two-stage design idea used in the CPS, like the Consumer Expenditures Surveys, the Survey of Income and Program Participation, the National Crime Victimization Survey, and the National Health Interview Survey. Some others are two-stage with selection of a list source followed by sampling from the lists like the Schools and Staffing Survey, the Private School Survey, and the Survey of Inmates of Local Jails. Still other are stratified samples from a sampled frame, such as the National Survey of College Graduates that has sampled from the Decennial Census Long Form, and the American Time Use Survey that samples from the CPS. In the early 1990s, The U.S. Census Bureau initiated the development of the use of continuous measurement as a possible replacement for the Decennial Census Long Form. Those efforts have since evolved into the current American Community Survey, which, starting 2010, will provide continual mid-decade estimates down to the block group level. The Census Bureau's goal for improving our sampling methodology to the present leads us to explore alternative sample designs.

3. Alternative survey sample design seminar series

The exploration into alternative methods of sampling began with an initial seminar series that was held at the U.S.

[1] Patrick E. Flanagan and Ruth Ann Killion, U.S. Census Bureau. E-mail: Patrick.e.flanagan@census.gov.

Census Bureau. It consisted of three seminar presentations of such methods covering the statistical bases of the methods and their limitations, especially when applied to the types of demographic surveys conducted by the U.S. Census Bureau. Each presentation also included discussant comments by Professor Jean Opsomer from Colorado State University. Three articles were then developed providing greater detail on each topic and a final discussant article covering the three subjects.

- On 26 September 2007, Professor Steven K. Thompson from Simon Fraser University gave a presentation on his research into network sampling, spatial sampling, and adaptive sampling.
- On 9 January 2008, Professor Sharon Lohr from Arizona State University gave a presentation on her research into sampling using overlapping frames.
- On 4 June 2008, Professor Yves Tillé from University of Neuchatel gave a presentation on his research into balanced sampling.

The articles resulting from this project that follow are:

"Adaptive network and spatial sampling," by Steven Thompson;

"Alternative survey sample designs: Sampling with multiple overlapping frames," by Sharon Lohr;

"Ten years of balanced sampling with the cube method: An appraisal," by Yves Tillé; and

"Innovations in survey sampling design: Discussion of three contributions presented at the U.S. Census Bureau," by Jean Opsomer.

4. Next steps

Following these three presentations, it was decided to conduct further research into these methods and their application to either existing U.S. Census Bureau Demographic surveys or to potential new surveys. There is already an urgent need for using multiple overlapping frames methods applied to the National Survey of College Graduates to deal with an old-cohort/new-cohort problem and a possible use of state hunting and fishing license registries as a second frame for the Fishing, Hunting, and Wildlife-Associated Recreation survey. We have plans to look at balanced sampling, particularly for selecting geographic primary sampling units. Lastly, the methods of adaptive sampling have the potential for us to accept surveys that we traditionally have not taken on, as well as providing a lower cost alternative for surveys that meet certain criteria.

5. Summary

This exploration into these three areas of alternative sample designs is just the beginning of our seminar series and of our intentions to explore methods to improve our demographic survey sample design methods. Future anticipated subjects include alternative listing methods, Kish's half-open interval approach to growth updates and coverage improvement, responsive survey designs, rejective sampling procedures, and model-assisted sampling.

Acknowledgements

This report is released to inform interested parties of research and to encourage discussion. The views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

References

Dedrick, C.L. (1938). *Census of unemployment* 1937: *Principle findings of the enumerative check census*. U.S. Bureau of the Census.

Duncan, J.W., and Shelton, W.C. (1978). *Revolution in United States Government Statistics* 1926 – 1976. U.S. Department of Commerce.

Frankel, L.R., and Stock, J.S. (1942). On the sample survey of unemployment. *Journal of the American Statistical Association*, 37, 77-80.

Stephan, F.F., Deming, W.E. and Hansen, M.H. (1940). The sampling procedure of the 1940 population census. *Journal of the American Statistical Association*, 35, 615-630.

Stephan, F.F. (1948). History of the uses of modern sampling procedures. *Journal of the American Statistical Association*, 43, 12-39.

U.S. Bureau of Labor Statistics and U.S. Census Bureau (2006). *Design and Methodology*: *Current Population Survey*.

Survey Methodology, December 2011
Vol. 37, No. 2, pp. 183-196
Statistics Canada, Catalogue No. 12-001-X

Adaptive network and spatial sampling

Steve Thompson [1]

Abstract

This paper describes recent developments in adaptive sampling strategies and introduces new variations on those strategies. Recent developments described included targeted random walk designs and adaptive web sampling. These designs are particularly suited for sampling in networks; for example, for finding a sample of people from a hidden human population by following social links from sample individuals to find additional members of the hidden population to add to the sample. Each of these designs can also be translated into spatial settings to produce flexible new spatial adaptive strategies for sampling unevenly distributed populations. Variations on these sampling strategies include versions in which the network or spatial links have unequal weights and are followed with unequal probabilities.

Key Words: Network sampling; Snowball sampling; Random walk; Markov chain; Adaptive web sampling.

1. Introduction

An adaptive sampling design is a procedure for selecting the sample in which the probabilities of selecting the set of sample units from the population depend on values of the variable of interest observed during the survey. In a spatial setting, adaptive sampling is exemplified by a survey in which, whenever a unit in the sample is observed to have an unusually high or otherwise interesting value of the variable of interest, nearby units may be added to the sample. In a network setting such as a socially networked human subpopulation, a link-tracing design may be used to adaptively follow social links from sample individuals to locate and add additional members of the subpopulation to the sample.

In spatial settings the development of adaptive designs has been motivated by such problems as estimating the abundance of rare, clustered plant and animal species, assessment of unevenly distributed environmental pollutants, and surveys of geographically clustered subpopulations of people. In network settings the development of adaptive network sampling designs has been motivated by problems in sampling people with rare diseases, sampling hidden populations such as those at high risk for HIV/AIDS or other epidemics, and sampling through computer and communications networks.

Zacks (1969) and Basu (1969) recognized that in most cases the optimal sampling would in principle be an adaptive one. With a Bayes model for the population, at any step part way through a sampling procedure, one can do as well or better than a conventional design by selecting the remaining sample to give the lowest mean square error conditional on the observed sample values so far. The overall mean square error is the expected value of the conditional mean square error. The underlying mathematical principle is that the integral of the minimum of a set of functions is smaller, or nor larger than, the minimum of the integrals. Results on optimal adaptive adaptive strategies are described and extended in Thompson and Seber (1996) and exemplified in Chao and Thompson (2001).

In spite of the early theoretical results and motivation from field surveys, the importance of adaptive designs was not widely recognized for several decades in either theory or practice. The practical importance of adaptive sampling strategies became evident as statistical thinking was brought to bear on problems in natural resource management and environmental protection. The development of adaptive link-tracing designs for reaching hidden human populations has attained strategic importance for such problems as understanding and alleviating the global HIV epidemic. In addition, new interest in adaptive sampling methods is being spurred by problems of expense and effort in social surveys of all types.

Adaptive designs such as those described in this paper often serve as high yield designs in that sample values of variables of interest tend to be higher on average than population means of the same variables. Although this is often a desired characteristic itself in studies of rare populations, simple sample data summaries such as sample means and sample proportions are generally not good estimates of population means or proportions. Instead, effective design-based and model-bases estimators of population quantities have been developed for use with adaptive designs.

With design-based estimators, properties such as unbiasedness or consistency depend solely on the way the sample is selected and not on assumptions about what the population may be like. Model-based estimators such as maximum likelihood or Bayes estimators on the other hand require use of a statistical model, usually involving unknown parameter values, describing the population of interest. Design-based estimators for adaptive designs are described in Thompson and Seber (1996), Thompson (2006a, b), and earlier papers.

[1] Steve Thompson, Simon Fraser University. E-mail: Thompson@stat.sfu.ca.

Basic results for model-based approaches to inference with adaptive designs were given in Thompson and Seber (1996), which showed that likelihood-based methods such as maximum likelihood and Bayes inference would be more effective than other model based approaches (for example, the linear unbiased prediction approach) with adaptive designs. Maximum likelihood estimation and the likelihood based approach more generally with link-tracing designs were described in Thompson and Frank (2000). Bayes estimation with link tracing designs was used in Chow and Thompson (2003). A method combining model and design based features was used in Felix-Medina and Thompson (2004). Bayes estimation using Markov chain Monte Carlo (MCMC) with adaptive web sampling designs is described in Kwanisai (2005, 2006).

2. Adaptive sampling in network settings

A population has network structure if there are links or relationships between any of the units in the population. Mathematically, such a population is described as a graph, consisting of a set of nodes and a set of edges or arcs between nodes. More generally, each relationship between a pair of nodes may have a weight denoting the strength of value associated with the relationship.

Human populations have an inherent network structure arising from social relationships. As will be noted later, spatial relationships also give a network structure to many populations. Network populations also arise in computing networks, communications, gene regulation and metabolic networks.

Network structure in populations is important for two reasons. First, the network relationships may be of interest in themselves to researchers. For example, with contagious disease epidemics it is important to know the nature and pattern of the social contacts through which the disease spreads. Second, the network structure can be used to help in obtaining a sample from a population that is otherwise difficult to sample. For example, in the study of hidden populations at risk for HIV/AIDS, including drug injectors, commercial sex workers, and others, often the only way in many cases to obtain a sample large enough for the study is to follow social links from initial sample individuals to find more members of the hidden population.

Most network sampling designs which follow links are inherently adaptive in that the link values used in the selection are variables of interest that are generally not known prior to the survey. Further, in some studies it may be of interest to follow links with higher probability from sample individuals with high values of variables associated with behavioral risk.

A class of designs called *multiplicity sampling* or simply *network sampling* was introduced my Birnbaum and Sirken (1965), along with design-unbiased estimators of population quantities. The approach was developed further by Sirken (1970, 1972a, b) and others and is described in Thompson (2002). In these designs the units on which observations are made are obtained by first selecting “selection units”, to which the observational units are linked. Motivation for these strategies came from problems in public health in which commonly used estimates were found to be biased because of the unequal numbers of such links. The simplest of the unbiased estimators in terms of computations was the “multiplicity estimator” which simply divided the observed value of a variable of interest measured on an observational unit by its “multiplicity”, the number of selection units to which it is linked. Horvitz-Thompson estimators for the strategy were also introduced. The following decades saw many variations on this strategy published in the statistics and substantive literatures.

In *snowball sampling* an initial sample of nodes is selected by some design such as simple random sampling, and every link out is followed to add connected nodes to the sample. This process is continued for a specified number of steps, or “waves”. More generally, a subsample such as a fixed number of links are followed at each wave. Frank (1971, 1977a, b, 1978a, b, 1979) framed the problem as one of sampling in graphs worked out design-based estimators for many cases of snowball designs including designs with unequal initial selection probabilities and estimators for population quantities such as totals and means of variables associated with nodes or individuals, as well as of population link quantities such as mean degree, where degree of a node is defined as the number of links out (or in) from that node. Frank and Snijders (1994) introduced a number of design-based and model-based estimators for one wave snowball designs motivated by the problem of estimating the number of injection drug users in a city.

In a *random walk* design an initial node is selected at random. From the links out from that node one link is selected at random and followed to add the connected node to the sample. This process is continued for a specified number of waves, with one unit selected at each wave. If the sampling is with replacement the design is a Markov chain, with the state of the chain at each step being the identity of the node selected at that step. Properties of such designs, cast as Markov chains in graphs, such as the limiting or stationary probabilities were examined in the statistics and probability literatures (Lovász 1993). Random walk designs were introduced into the social network literature by Klovdahl (1989) with the motivation of reaching into a hidden human population farther away from the initial sample than possible with the same sample size using a snowball design. In the computing science literature, Brin and Page (1998) used the concept of a stationary distribution of a random walk in a graph in developing a search engine and web page ranking algorithm, evoking the metaphor of a "random surfer" to describe the process of a random walk following hyper-links from web page to web page.

Heckathorn (1997, 2002) and Salganik and Heckathorn (2004) described a sampling methodology referred to as "respondent driven sampling" in which members of a hidden population were motivated to recruit other members of the population into the sample using a system of coupons. A simple estimator of population totals and means, in which each observation is weighted by the reciprocal of that person's degree, was used with these designs based on the limiting distribution of a with-replacement random walk in a network having symmetric links and a single connected component. The coupon-based methodologies developed with these designs have proven to be highly effective in recruiting samples of substantial size from hidden populations in a number of settings.

The notational setup for sampling in networks follows. There is a population of units or nodes labeled 1, 2, ..., N with associated variables of interest $y_1, y_2, ..., y_N$. Associated with each *pair* of nodes (i, j) is a link-indicator or weight, so that the collection $\{w_{ij}; i, j = 1, ..., N\}$ are variables of interest associated with pairs of nodes.

In the network context a sample s is a subset $s^{(1)}$ of nodes and a subset $s^{(2)}$ of the pairs of nodes, that is, $s = (s^{(1)}, s^{(2)})$. Thus the sample consists of a sample of nodes, on which node variables y of interest are recorded, and a sample of pairs of nodes, for which the values of relationship variables w are recorded.

Figure 1 shows a network-structured population which will be used to illustrate some of the network sampling designs described in this paper. In terms of a human population with social network structure, the red or dark colored nodes could represent individuals with high values of variables of interest, for example indicating a risk-related behavior such as injecting drug use. The light colored or yellow nodes would represent the individuals without the high-risk characteristic of interest. The links between individuals would represent social relationships such as having meals together, drug-using relationships, or sexual contacts.

Figure 2 shows an initial simple random sample of five nodes selected from the network-structured population. A one-wave snowball sample selected by following every link out from the initial sample is shown in Figure 3, and a two-wave snowball sample from the same initial sample is shown in Figure 4. Note that with a fixed number of waves, a snowball sample can grow very fast.

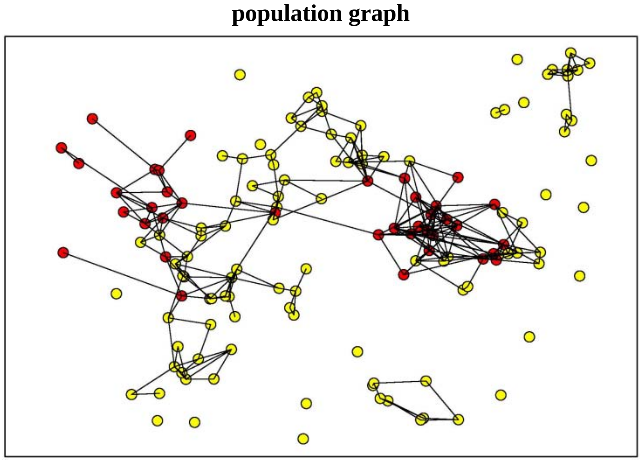


Figure 1 A population with network structure

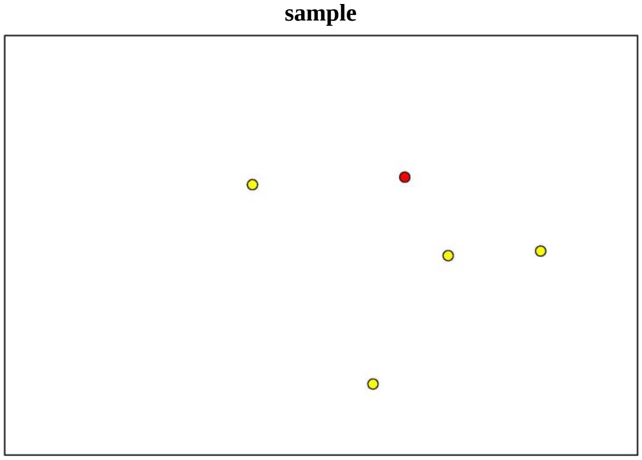


Figure 2 A random sample of nodes

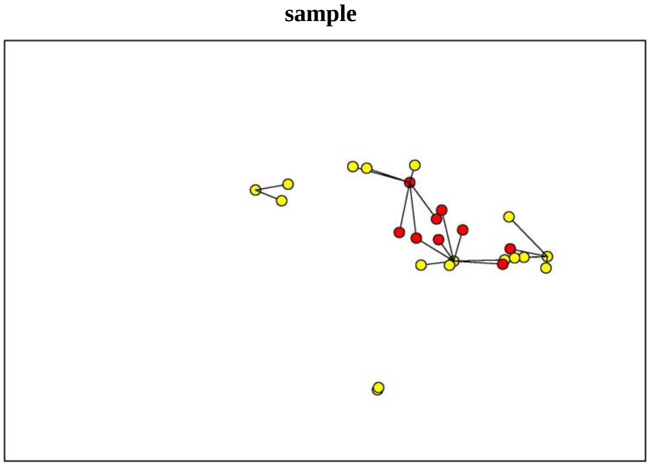


Figure 3 One-wave snowball sample

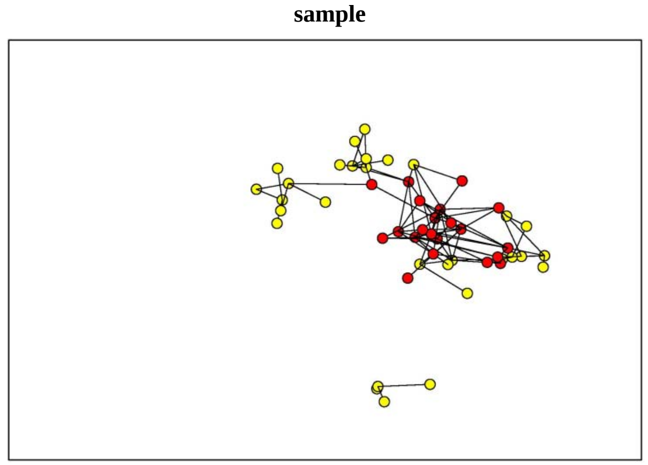


Figure 4 Two-wave snowball sample

With a snowball sampling designs and many other link-tracing designs, sample data summaries such as a sample mean or sample proportions are not good estimators of the analogous population characteristics. The reason is that under the design different units have different probabilities of selection, dependent on the population link structure. Figure 5 shows the population with the size of each node proportional to the probability of selecting that node. Since high-risk individuals tend to have more links hence higher probabilities of inclusion in the sample, the sample mean would tend to overestimate the population mean. In the same way, the average degree of such a sample would tend to overestimate the mean degree of the population network.

With the one-wave snowball design in a setting with symmetric links the inclusion probabilities for sample nodes can be easily calculated as proportional to the node degrees. With asymmetric links or with snowball designs of more than one wave it is not in general possible to calculate node inclusion probabilities from the sample data. Methods for calculating design-unbiased estimators of population node and link characteristics with such designs are described in the section on adaptive web sampling later in this paper.

Figure 6 shows a snowball sample from this same network population starting with one randomly selected unit. Since the population consists of more than a single connected component a strict random walk design would be stuck in whatever component it started in. It is therefore desirable to provide in the design some small probability at each step of selecting the next unit by simple random sampling or some other conventional design, or at least allowing a random jump whenever a walk is found to be stuck in a component.

Figure 7 shows the stationary selection probabilities for the random walk through the network shown. Although these probabilities in this population are not simply proportional to node degrees it can be seen that nodes with high degree do tend to have high selection probabilities. Also, since high risk individuals in this population tend to have high selection probability under this design, sample summaries such as sample mean and sample proportion are not unbiased estimators of population means and proportions. For unbiased estimates the methods of later sections of this paper would have to be used.

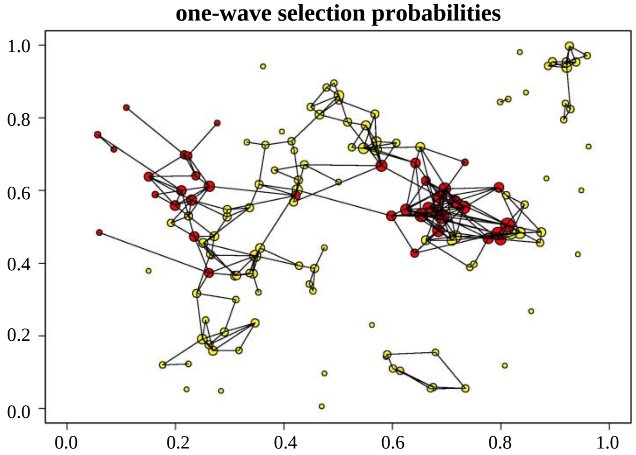


Figure 5 One-wave snowball sample selection probabilities

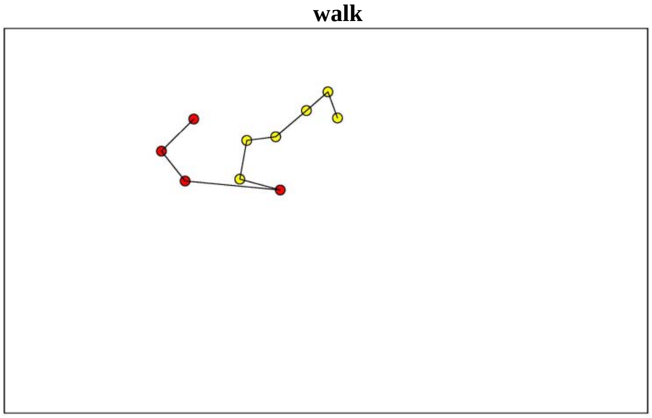


Figure 6 A random walk sample from the same population

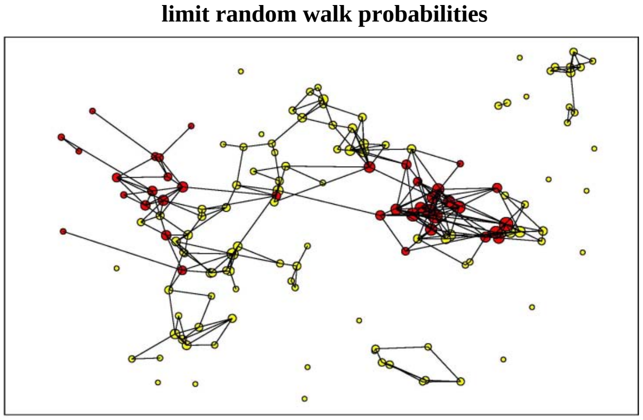


Figure 7 Random walk limit selection probabilities

2.1 Targeted random walk designs

One of the early motivations for using random walk designs with hidden populations was to penetrate deeper into the population, that is, farther from the initial sample and thereby obtain a more representative sample of the population. When the probabilities of selecting a given person by such a method are calculated either step by step or in their stationary limit, they are not in general equal but depend on the link and degree structure. With the motivation first to find a method for selecting a sample through a network such that the stationary probabilities would be the same for each person or node, uniform and targeted random walk sampling designs were developed (Thompson 2006a). An additional motivation was to find a more flexible and adaptable way to sample through a network.

Since a random walk with replacement through a graph or network is a Markov chain, ideas of Markov chain Monte Carlo can be applied to produce a different Markov chain having desired stationary probabilities. At each step of the sampling the state of the chain is the current node added to the sample. The stationary probabilities of the chain correspond to the stationary selection probabilities for each person or node. With a targeted walk design the random walk design is tweaked at each step, based on out-degree of each node, to obtain a design with specified limiting selection probabilities.

Suppose that at some step in the sampling person i is the last person who has been added to the sample. Using a random walk procedure we randomly select one of the links out from that person, and that link leads to person $j,$ who is now our tentative selection. A screening interview reveals that person j has more links out than person $i,$ so that the conditional probability of going from i to j as we just did is larger than the conditional probability in the reverse direction, since the transition probabilities are related to the reciprocal of the number of links out. Therefore we calculate a probability less than one and accept person j into the sample only with that probability. If our tentative selection is not accepted we independently again choose a link out from person $i.$ The probability of acceptance of the candidate link is based on the Hastings (1970) generalization of the Metropolis algorithm. The acceptance probability depends on the desired target selection probabilities, the number of links out from the current node and the candidate node, and the probability of going in either direction with a random jump if that is part of the design (Thompson 2006a).

Note that the method depends only on links out, which can usually be determined for sample members, whereas links in to sample individuals usually can not be determined. Therefore the method applies to directional as well as symmetric networks.

A uniform walk design is the special case in which the targeted stationary selection probabilities are all equal. A targeted random walk design could be used for example to obtain a sample from a hidden population in which an individual with a certain high-risk behavior would have selection probability twice that of an individual without the behavioral characteristic.

It is the sample of accepted people or nodes that has the desired stationary selection probabilities. If the tentative selections had been interviewed thoroughly also, not only the screening interview about out-degree, then in principle the estimates from the accepted sample could be improved using the Rao-Blackwell method (Casella and Robert 1996). That would involve calculation of the probabilities of getting the same data with different accept-reject results and in different orders of selection. With each of the different accept scenarios the estimate would be computed using the accepted set and each value weighted by the ordered selection and acceptance probabilities. In most cases there are too many combinations for exact calculation, and a more practical approach would be the Markov chain resampling method at the inference stage described in a later section of this paper. It is not clear that in practice it would be desirable to compute the improved estimators using the data since full interviews rather than screening interviews would be required for those not initially accepted, the computations for the improvement are potentially demanding, and the calculation depends on knowing the selection probabilities for the initial sample, which is not needed for the simple estimators.

With a targeted walk design in which the target stationary selection probability π_i of node i is proportional to $c_i,$ an asymptotically consistent estimator, based on the limiting probabilities, is provided by the generalized ratio estimator

$$\hat{\mu} = \frac{\sum_{s_a} y_i / c_i}{\sum_{s_a} 1 / c_i}$$

where y_i is the value of the variable of interest for the i^{th} node and s_a is the sample of selected nodes. In this type of estimator the relative values of target probabilities need be specified since the proportionality constant cancels out.

Note that a straight Horvitz-Thompson or Hansen-Hurwitz estimator can not be used because the proportionality constant in the inclusion probabilities is unknown, whereas in the generalized ratio estimator it cancels out. Again the limiting probabilities on which the estimator is based hold exactly for the with-replacement design. For the without-replacement variation, the properties of the targeted strategies were fairly closely approximated by the with-replacement properties in the empirical comparisons (Thompson 2006a).

2.1.1 Designs using weighted links

Many studies of socially networked populations conceptualize the network as having nodes (people) and lines or arrows representing the links or relationships between people. The network is characterized by an incidence matrix of 0s and 1s indicating when there is a link from node (row) i to node (column) j. In many real situations, however, more than one type of link may be of interest and links may have different weights representing differing strengths of a relationship. For example, in studies of risk behaviors and interventions in relation the the HIV epidemic, two types of links of high interest are sexual relationships and drug injecting relationships. Other social relationships, such as friendships and living arrangements, may also be of interest to investigators and may be useful in finding members of the population. These types of relationships may have weights corresponding to frequency of encounters, geographic proximity, or other measures of strength.

In the basic form of weighted link designs we consider, in which one link from the most recently selected person is selected from the links out from that person, the selection is made with probability proportional to link weight. More generally, the selection could be made based on that weight but not necessarily proportional to it. However, we could then redefine the weight to be proportional to the probability we have under the design of following that weight, so that the following result would still apply.

The following derivation shows that under suitable conditions the stationary selection probability for each person with such a design is proportional to the sum of the link weights out from that person. The result applies for a population in which it is possible to reach any one person from another following some path in which each link has weight greater than zero. That is, the population has a single component.

For such a condition to hold it is advantageous to have at least some probability of following common but weak links. For example, a study of a sexually transmissible epidemic may want to focus with high probability on sexual links. But sexual links do not connect the population into a single component. Therefore, some smaller probability is allowed in the design for following friendship or geographic links, which represent weaker relationships between people and are of less inherent interest to investigators but serve to connect the population. Thus, the combination of different types of links in this situation turn the population into a single component for purposes of the design.

2.1.2 Stationary distribution of weighted link Markov chain design

In this section we derive the stationary distribution of a weighted link design in a single component situation. Keep in mind that we may create the single component property through innovative use of geographic links in combination with social links.

Let w_{ij} be the weight of a link between node i and node j, and assume that these links are symmetric, so that $w_{ij} = w_{ji}$. Consider a random walk design, with replacement, in which the transition probability to node j, given the walk is at node i, is proportional to w_{ij}. That is, one link is selected out from node i with probability proportional to weight. The transition probability is thus $P_{ij} = w_{ij} / w_{i\cdot}$. The sum $w_{i\cdot} = \sum_j w_{ij}$ is the total weight out from node i, generalizing the concept of degree with equally weighted nodes.

Suppose the graph has only a single component, that is, any node in the graph can be reached from any other node by a path in which every link has positive weight. Then the stationary probability for node i is proportional to $w_{i\cdot}$.

Suppose that the probability that the walk is at node i at time t is $\pi_i = w_{i\cdot} / w_{\cdot\cdot}$, for $i = 1, ..., N$, where $w_{\cdot\cdot} = \sum_i \sum_j w_{ij}$, the total of all the weights. Then the probability that the process is at node i at time $t+1$ is $\sum_j \pi_j P_{ji}$ by the law of total probability. In terms of the link weights, this sum is $\sum_j (w_{j\cdot} / w_{\cdot\cdot})(w_{ji} / w_{j\cdot}) = \sum_j w_{ji} / w_{\cdot\cdot}$. Because of the symmetry of the weighted links, this becomes $w_{i\cdot} / w_{\cdot\cdot}$, so that if node i has this probability at time t it has the same probability at time $t+1$, so that these are the stationary probabilities of the process. By induction, once the process reaches it's stationary distribution it remains in it for every step thereafter. In practice, especially with small sample sizes or with different design variations, the stationary distribution serves as an approximation to the exact distribution.

If the weights are not symmetric, the selection probabilities of the random walk design will still approach a stationary distribution provided there is only a single component or, if not, that the design incorporates random jumps. However, with the directional weighted links, the stationary distribution is no longer of the simple form that can be calculated from sample data.

2.1.3 Different uses of weighted link designs

Variations of weighted link designs could prove useful in situations of the following types.

(1) Designs using general weights of links, on a continuous or discrete scale, representing strength or importance of relationships and probability of following them.

(2) Situations with two types of links, represented by two weights, such as social networks with strong and weak relationship links, or an HIV-at-risk study focusing on both sexual contacts and drug using relationships.

(3) Survey settings in which links represent the geographic or "random jump" part of the design, or the seed design. For example, all people within a given geographic stratum are linked by a geographic link, or all the people who visit any of the venues on an ethnographic map are thereby linked.

(4) In a situation where a sampling frame exists but the frame covers only part of the population, all units within the frame can be considered to be connected by a "frame link". Venue-based sampling typically forms one example of this type of situation.

(5) Using a variation on the sampling design as a model for the way a virus or other infectious agent "samples" people in a population. A type of weighted link design could be developed as a model for the spread of an infectious disease, finding the different importance of different links. For influenza, the relative importance of air transported droplets (sneezing, coughing) versus indirect contact through solid objects (door knobs, money). For HIV, the relative importance of different types of sexual contacts and unsafe injections, whether for illegal drugs or unsanitary medical injections especially in third world countries. The disease transmission in a simulation has a slightly different protocol than the implemented designs, in that instead of thinking of one new link selected at each selection time step, there could be anywhere from zero to a high number of transmissions in a time step.

2.1.4 Properties of weighted link designs and associated population graphs

Suppose the relationships in the population are assigned weights, with the weight w_{ij} denoting the strength of the relationships from node i to node $j.$ And suppose we use a link tracing design of the walk type in which the transition probability is

$$P_{ij} = \frac{w_{ij}}{w_{i\cdot}}$$

where $w_{i\cdot} = \sum_{j=1}^{N} w_{ij}$. This is the conditional probability of selecting node j as the next sample unit, given the most recently selected unit is node i. The walk design is a Markov chain on a graph, in which the graph has weighted links.

We will next consider the question in the other direction of when a Markov chain can be represented by a design of this sort on a graph with weighted links. Given a Markov chain specified by a matrix of transition probabilities P_{ij}, we can always represent it as a walk design of this type on a graph with weighted links so long as the links satisfy the first of the following properties:

(1) $w_{ij} = P_{ij} w_{i\cdot}$, where the row weight totals are arbitrarily chosen.

Next consider imposing some property on the weight row totals to make them unique. For example:

(2a) If the $w_{i\cdot}$ weight row totals are chosen to be all equal to a constant such as one, then the link weights represent the conditional transition probabilities given the process is at the node at which they originate.

(2b) If the $w_{i\cdot}$ weight row totals are proportional to the stationary probabilities π_i of the Markov chain for each node i, or equal to them, then the weights represent "flows" of the Markov chain, that is, the unconditional probabilities of transitions along the links:

$$w_{ij} = P_{ij}\,\pi_i.$$

In the practical situations for which we are trying to find appropriate models and designs, the weights may be at least partially given by the natural circumstances of the situation. For example the weight w_{ij} may represent the presence or absence of a link from person i to person j, or the number of transactions of a certain type in a given time period from i to j. In that case, condition (2a) above would not in general be satisfied and condition (2b) would be satisfied only if all the weights were symmetric, that is, if $w_{ij} = w_{ji}$ for all i and j.

In particular, if some or all of the weights are asymmetric, with $w_{ij} \neq w_{ji}$, then (2a) would not usually be satisfied and it would not be possible to arbitrarily choose weights to impose the condition because typically the stationary probabilities would not be known and could not be calculated from the sample data. However, although the row totals $w_{i\cdot}$ could not be arbitrarily imposed, they can be known for units in the sample since they are simply the total weight out from each unit.

2.2 Adaptive web sampling

Targeted random walk designs provide considerable flexibility and control not offered by regular random walks. The use of weighted links with these designs extends that flexibility farther. This flexibility is still constrained, however, by the restriction that the selection of the next link to follow can depend only on the most recently selected node in the sample. The incentive for developing the next set of designs was to remove this restriction and greatly expand the scope for flexibility and control in the available strategies.

In an adaptive web sampling design (Thompson 2006b) an initial sample of one or more unit/node is selected by simple random sampling or other conventional design. From then on, at each step in the sampling there is an active set consisting of the sample selected so far or some subset of it.

In the simplest case, one link is selected from the links out from this set. Sampling continues in this fashion until the desired sample size or some other stopping criteria has been satisfied. Some small probability is allowed, however, that the next node is selected at random, or by some other conventional design, from the entire population. The designs can be done with or without replacement.

More generally a set of links can be selected at each step. Also the links at each step can be selected by a design more complicated than simple random sampling. The selection probabilities can be dependent on node or link characteristics and can be varying over time.

The basic idea of an adaptive web sampling design is shown in the next set of figures. In Figure 8, an initial sample of two nodes has been selected by random sampling without replacement. At the next step a link may be chosen out at random from either of the initial nodes to add a new node to the sample, as shown in Figure 9. The next node is selected by following one of the links out from the current sample. With a random walk a link would need to be followed from the last node selected, but with adaptive web sampling any eligible link out from the current sample (active set) may be followed. Note the next selection, shown in Figure 10, is not via a link from the most recently selected node, but from a previous one. As sampling progresses it is free to branch out flexibly in different directions as well as select new nodes at random from the population (Figure 11). The design can be stopped at a specified sample size or some other criteria. In the design shown in the figures, links out from the current sample were not selected completely at random but with higher probability given to following links from high-risk individuals, represented by dark or red nodes. Further, the design shown allowed a 0.1 probability of selecting the new node at random at any step instead of following a link.

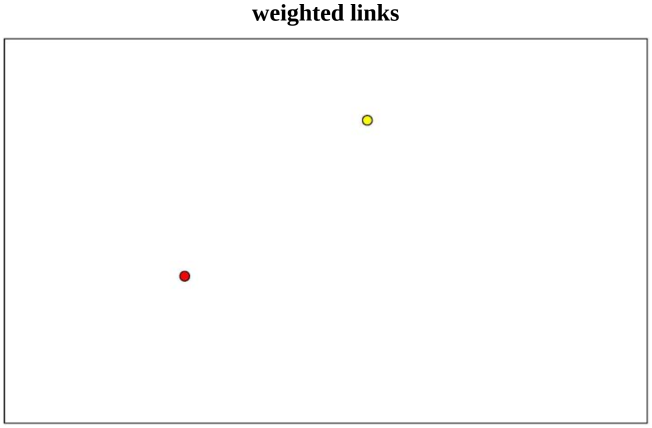


Figure 8 The first two nodes selected at random

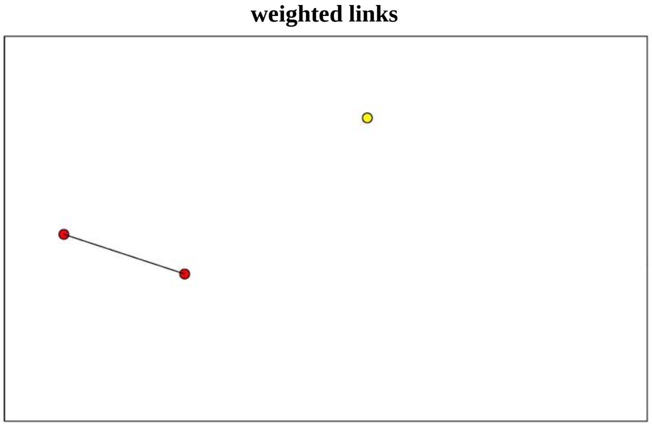


Figure 9 The next node is selected by following one of the links out from the current sample

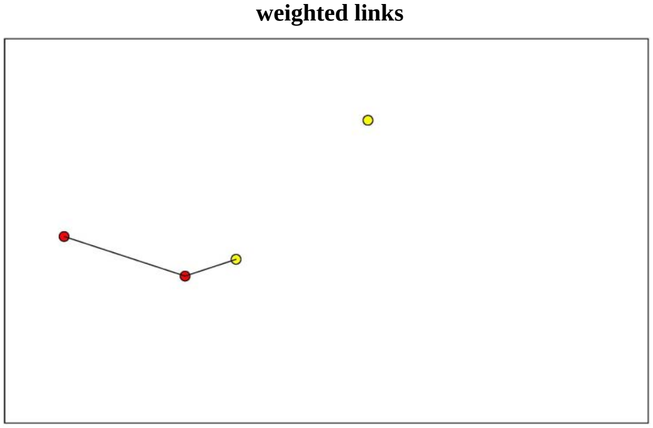


Figure 10 Note the next selection is not via a link from the last-selected node, but from a previous one

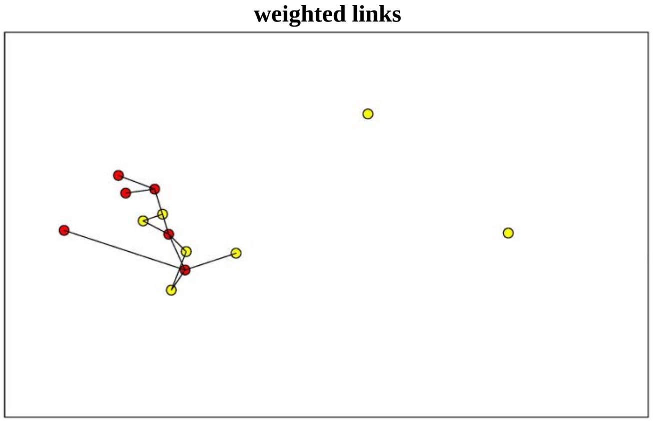


Figure 11 As sampling progresses it is free to branch out flexibly in different directions as well as select new nodes at random from the population

2.2.1 Inference methods

Design-unbiased and design-consistent estimation methods for use with adaptive web sampling designs are described in Thompson (2006b). Bayes model-based estimation methods for use with adaptive web sampling are described in Kwanasai (2005).

The design-based estimators are constructed by starting with some relatively easy to compute estimator that depends on the order of selection of the sample. This initial estimator is then improved using the Rao-Blackwell method, that is by obtaining the expected value of the initial estimator conditional on the minimal sufficient statistic.

2.3 Estimator based on initial sample mean

Suppose $\hat{\mu}_0$ is an unbiased estimator of the population mean that depends on the order in which the sample is selected. If the initial sample of nodes has been selected by simple random sampling, one example of an unbiased initial estimator that depends on order is the initial sample mean. The improved estimator has the form

$$\hat{\mu} = \mathrm{E}(\hat{\mu}_0 \mid d_r) = \sum_{\{\mathbf{s}:r(\mathbf{s})=s\}} \hat{\mu}_0(\mathbf{s})\, p(\mathbf{s} \mid d_r).$$

Here $\mathbf{s}$ denotes the sample in order of selection, r is the reduction function that reduces the ordered sample to s, the unordered sample of the minimal sufficient statistic. The reduced data d_r consists of the unordered sample together with the associated values of the variables of interest. The improved estimator $\hat{\mu}$ is the expected value of the initial estimator over all $n!$ reorderings of the sample data. In calculating the expectation, each of the reorderings is weighted by the selection probability $p(s \mid d_r)$.

Other initial estimators used with adaptive web sampling utilize the entire sample data but depend on order and are based on using the conditional probabilities of selecting each new unit in sequence given the previously selected units. Four types of design-based estimators for use with adaptive web sampling are given in Thompson (2006b).

Computation of the improved estimator $\hat{\mu}$ and its variance estimators under various adaptive web designs involves enumerating the reorderings of the sample selection sequence. For each reordering, the probability of that ordering under the design is computed, along with the values of the estimators and variance estimators. Direct calculation is fast and efficient up to sample sizes of ten or so, which involve no more than a few million permutations to be enumerated. For larger sample sizes, the numbers of permutations or combinations of potential selection sequences in the conditional sample space become prohibitively large for the exact, enumerative calculation. For this reason, a Markov chain resampling approach was used in Thompson (2006b) for computing the improved estimators.

The resampling procedure is as follows. The object is to obtain a Markov chain $x_0, x_1, x_2, \ldots$ having stationary distribution $p(x \mid d_r)$. Here x_k denotes an entire reordering of the sample at step k of the chain. Suppose that at step $k-1$ the value is $x_{k-1} = j$, so that h denotes the current permutation of the sample data in the chain. A tentative or candidate permutation c_k is produced by applying the original sampling design, with sample size n, to the data as if the sample comprised the whole population, that is, as if $N = n$. This resampling distribution, denoted p_c differs from, but has some similarity to, the actual sampling design p. The desired conditional distribution $p(x \mid d_r)$ is proportional to the unconditional distribution $p(x)$ under the original design applied to the whole population.

Let

$$\alpha = \min\left\{\frac{p(c_k)}{p(x_{k-1})}\frac{p_c(x_{k-1})}{p_c(c_k)},\ 1\right\}.$$

With probability α, t_k is accepted and $x_k = c_k$, while with probability $1-\alpha$, c_k is rejected and $x_k = x_{k-1}$.

This procedure produces a Markov chain $x_0, x_1, x_2, \ldots$ having the desired stationary distribution $p(x \mid d_r)$. The chain is started with the original sample $\mathbf{s}$ in the order actually selected. Given any value of the minimal sufficient statistic d_r, the chain is thus started in its stationary distribution and so remains in its stationary distribution step by step.

Suppose that n_r resampled permutations are selected by this process and let $\hat{\mu}_{0h}$ denote the value of the initial estimator for the h^{th} permutation. An enumerative estimator of the form $\hat{\mu} = \mathrm{E}(\hat{\mu}_0 \mid d_r)$ is replaced by the resampling estimator

$$\tilde{\mu} = \frac{1}{n_r}\sum_{h=0}^{n_r-1} \hat{\mu}_{0h}.$$

Bayes model-based inference with adaptive web sampling designs also requires the use of Markov chain Monte Carlo (MCMC) methods except in certain fairly simple design situations (Chow and Thompson 2003) where explicit Bayes posterior distribution, estimators, and intervals can be obtained. More generally the MCMC sequence involves at each step updating of model parameter estimates and, in a data augmentation procedure, obtaining a complete realization of the population network and its values from the predictive posterior distribution conditional on the observed data (Kwanisai 2005, 2006). The resulting Markov chain sequence of complete population realizations provides the flexibility to make inference about many types of population characteristics.

2.4 Modification of adaptive web sampling procedures

Adaptive web sampling designs are a generalization of random walk designs. The more general designs do not have the exact stationary distribution properties of walk designs, since more than one link may be followed from any node, links may be followed from sample nodes other than the most recently selected one, and the sampling may be done without replacement. However, the stationary distribution properties of a random walk or other Markov chain design may serve as a guide to approximate properties one might expect from a similar adaptive web sampling design.

During the sampling, at the time of the t^{th} unit selection in the k^{th} wave, let $w_{a_{kt}+}$ be the total number of links out, or the total of the weight values, from the active set a_k to units not in the current sample s_{ckt}. That is, $w_{a_{kt}+} = \sum_{\{i \in a_k,\, j \in \bar{s}_{ckt}\}} w_{ij}$. When w is an indicator variable, $w_{a_{kt}+}$ is the total of the net out-degrees of the individual units in the active set a_k, where net out-degree is the out-degree of a unit minus the number of its links to other units already in the current sample.

For each unit i in the sample, the variable of interest y_i and the out-degree (or out-weight) w_{i+} are recorded. In addition, for each pair of units (i, j) for which both i and j are in the sample, the values of the link variables w_{ij} and w_{ji} are observed.

Consider as a candidate for the t^{th} selection in the k^{th} wave a unit i not in the current sample, so $i \notin s_{ckt}$. Suppose the current active set a_k contains one or more units having links or positive weights out to unit i, and let $w_{a_k i} = \sum_{j \in a_k} w_{ij}$ denote their total. The probability that unit i is the next unit selected is

$$q_{kti} = b \frac{w_{a_k i}}{w_{a_{kt}+}} + (1 - b) \frac{1}{(N - n_{s_{ckt}})}$$

where b is between 0 and 1. If there are no links at all out from the current active set, then

$$q_{kti} = \frac{1}{(N - n_{ckt})}.$$

Thus, with probability b link-tracing is done, and one of the links out from the current active set is selected at random, or with probability proportional to its weight, and the node to which it leads is added to the sample, while with probability $1 - b$ the new sample unit is selected completely at random from the units not already selected. However, if there are no links or positive weights out from the active set to any unsampled units, then the next unit is selected from the collection of unsampled units.

Basic adaptive web sampling can be generalized to use weighted links. If the relationship variable w consists of weights, instead of having just 0 or 1 values, then the link-based selection can depend on these weights. For example, link weights can be defined in relation to the y value of an originating node or as a distance measure to the connected node, so that links are followed with higher probability from nodes with higher values or with lower probability to distant nodes. Then a link from the active set can be selected with probability proportional to link weight, or with some other selection probability $p(i \mid s_{ckt}, a_k, y_{a_k}, w_{a_k})$ depending on variables of interest only through the active set. For example, a link out could be selected at random from the links with w_{ij} greater than some constant, or y_i greater than some constant. The selection probability when links are not followed does not have to be uniform over the units not in the current sample, but can be a more general design $p(i \mid s_{ckt})$ such as selecting with probability related to an auxiliary variable or from a spatially defined distribution.

With weighted links w represents a possibly continuous link weight variable and the probability that unit i is the next unit selected is

$$q_{kti} = b\mathrm{p}(i \mid s_{\mathrm{ckt}}, a_{\mathrm{k}}, y_{a_{\mathrm{k}}}, w_{a_{\mathrm{k}}}) + (1 - b)\ \mathrm{p}(i \mid s_{\mathrm{ckt}}).$$

If there are no links or positive weights from a_k to i, then

$$q_{kti} = \mathrm{p}(i \mid s_{\mathrm{ckt}}).$$

Once unit i has been selected, it is possible to add an accept/reject step for deciding whether to include it in the active set, for example, accepting with higher probability if unit i has a high value or high degree.

In the design the constant b itself can also be replaced by a probability $b(k, t, a_k, y_{a_k}, w_{a_k})$ depending on values related to nodes and links in the active set or changing as sample selection progresses. For example, if the values of the units in a_k are particularly high, we could increase the probability of following links. As for dependence of b on (k, t), the use of an initial conventional sample of size $n_0 > 1$ may be viewed as serving to obtain some information from basic coverage of the population before adaptive sampling is allowed to commence.

3. Spatial adaptive web sampling

Adaptive sampling designs such as adaptive cluster sampling (Thompson 1990) were developed in response to the need for more effective strategies for sampling spatially uneven populations, particularly those having a rare, clustered geographic distribution. Most populations having a network structure also have an inherent geographic or spatial structure. For example, human populations have social network structure but are also distributed in space. Of particular interest from the sampling design point of view, spatial structures can be characterized with graph or network structures. For example, neighborhood relationships based on geographic proximity can be recast in the form of lattice-type graphs. In this way, network designs such as those described in the previous section can be applied to solve spatial sampling problems.

In this section the use of adaptive web sampling designs to sample a spatially uneven population will be described.

These designs could be viewed as a generalization of adaptive cluster sampling. In this view, adaptive cluster sampling would be a special case in which every link is followed until there are no more links out from the current sample. The adaptive web sampling class of designs offers more flexibility and control, however, and is potentially more efficient to use for many spatial populations.

With adaptive cluster sampling the constraint to continue to sample until all neighbors of all units satisfying the condition were included meant that overall sample size was not controlled in advance and was rather stringent when some networks were unusually large. Adaptive web sampling in the spatial context solves this problem since sample size can be fixed in advance. In terms of its network recasting, the simple unbiased estimators of adaptive cluster sampling use data only from the strongly connected components that the initial sample intersects. Rao-Blackwell improvements based on those estimators can use in addition data from the weakly connected extensions of those components. The familiar edge units of spatial adaptive cluster sampling are a special case of such weakly connected extensions of strongly connected components.

Figure 12 depicts a study region with a spatial clustered population as may be encountered in ecological, epidemiological, and social demographic surveys. In one form of adaptive spatial designs the neighborhood of a unit is defined as the set of immediately adjacent units, and neighboring units are added to the sample when the value of a sample unit is high or meets some other criterion. In Figure 13 the spatial population has been recast as a directed graph. The square spatial units are redrawn as nodes in a graph, and whenever the number of objects in a unit exceeds zero, arrows representing graph links are drawn from that node to neighboring nodes. Nodes representing units with nonzero values are colored dark (red). Figure 14 shows a random sample of nodes to be used as the initial sample of an adaptive web design. The adaptive web sampling continues until the targeted final sample size of 20 units is obtained in Figure 15. The sample is recast in the spatial setting in Figure 16. Unlike adaptive cluster sampling, it was not necessary to continue sampling until every unit in a sampled connected component is included. Further, the small probability of a random jump keeps the design from being stuck in any connected component.

A glimpse of the immense flexibility offered with the adaptive web sampling designs in the spatial setting is shown in Figure 17. In the top row a spatial population is recast as a graph, though the directions of the links are not shown. The bottom row shows samples from two variations of adaptive web sampling. On the left, sixteen initial units have been selected independently at random. From each, an adaptive web sampling procedure is carried out to a sample size of five units. With this design, the sample is spread throughout the study region while also reaching into components. In the design on the right a single initial unit is selected at random and adaptive web sampling continues to a total of 80 units. The 0.1 probability of selecting the next unit at random at any step prevents the design from being stuck in any one component. With this design the main components or aggregations get very thorough, though not exhaustive, coverage.

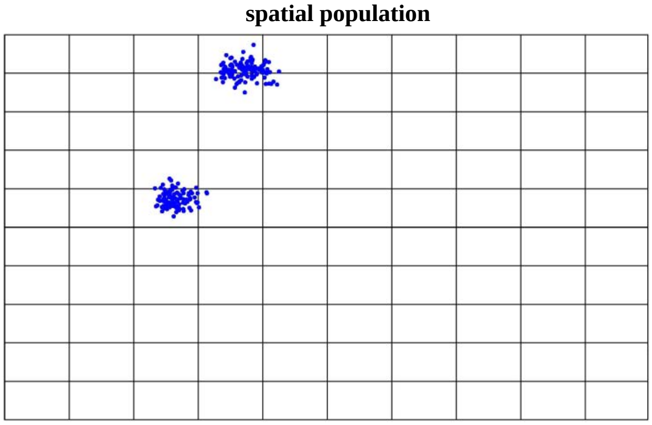


Figure 12 A spatially clustered population

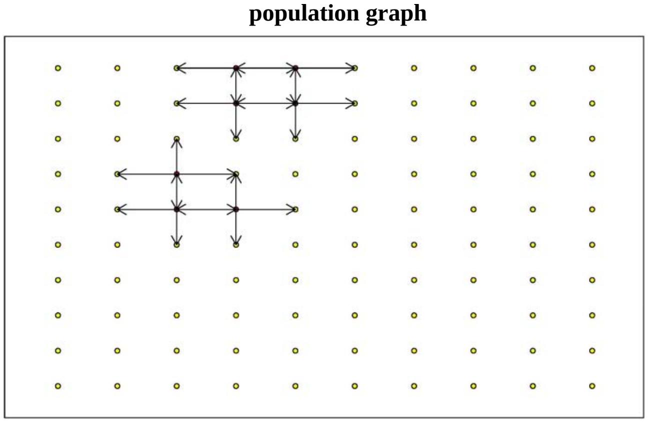


Figure 13 A network representation of relevant neighborhood relationships in the spatial population

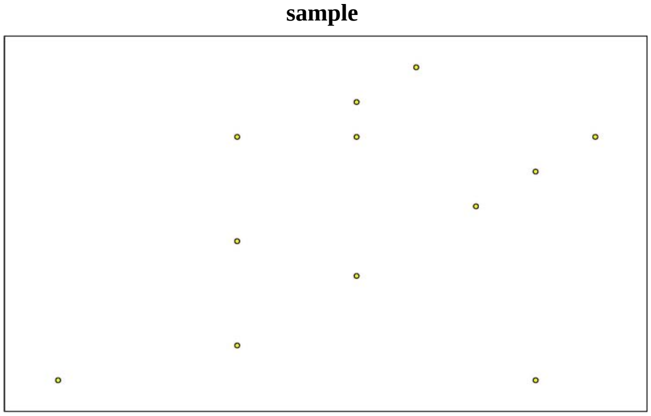


Figure 14 An initial random sample of spatial units

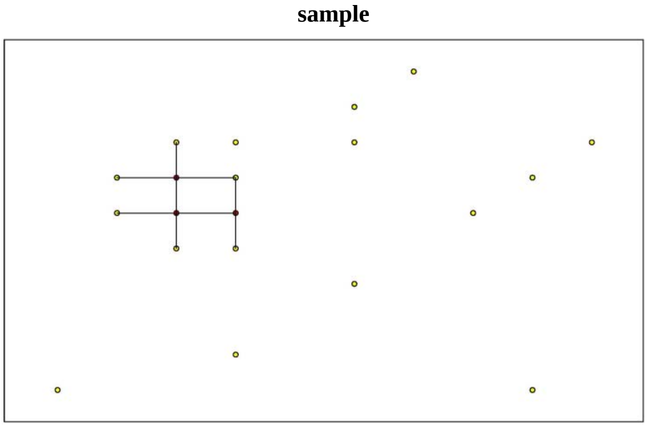


Figure 15 Adaptive web sample of 20 units starting from the initial sample of the previous figure

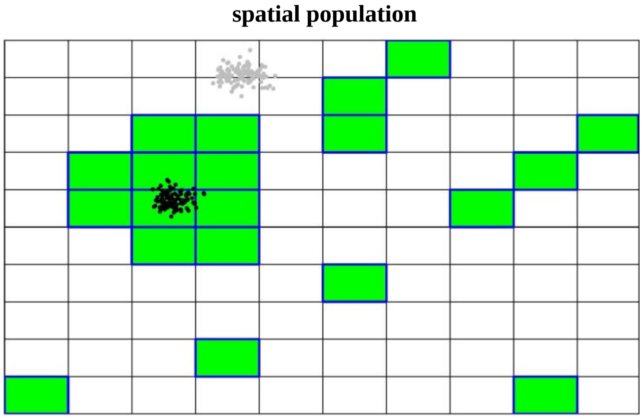


Figure 16 Spatial representation of the adaptive web sample

3.1 Spatial designs with weighted links

For selecting spatial samples, link weights can be defined as a function of the distance between sites. For example, for increased sample the function would give larger weight to sites at close distance. On the other hand, for space filling purposes sites at larger distance could have larger weight. A network sampling design in such a setting, with link weights defined solely on the basis of distance, would not in general be adaptive. That is because the spatial frame would enable a link-tracing design to select the entire sample of sites before going in the field to make any observations.

More generally though link weights can be defined as a function of both weights and observed values. For a unit in the sample having a high observed value of the variable of interest, the function could give higher weight at distances close to that site and smaller weight to distance sites. For a unit having a low value of the variable of interest the weight function could have a more uniform shape.

Random walk designs in particular are straightforward to carry out in spatial settings with links weights dependent only on distance. That is because at any point in the sampling the selection of the next site depends only on the most recently selected site, so that only one weight function needs to be considered. With more general designs such as adaptive web sampling the use of link weight functions dependent on both distance and value opens up very wide flexibility in the possibilities available for adaptive strategies.

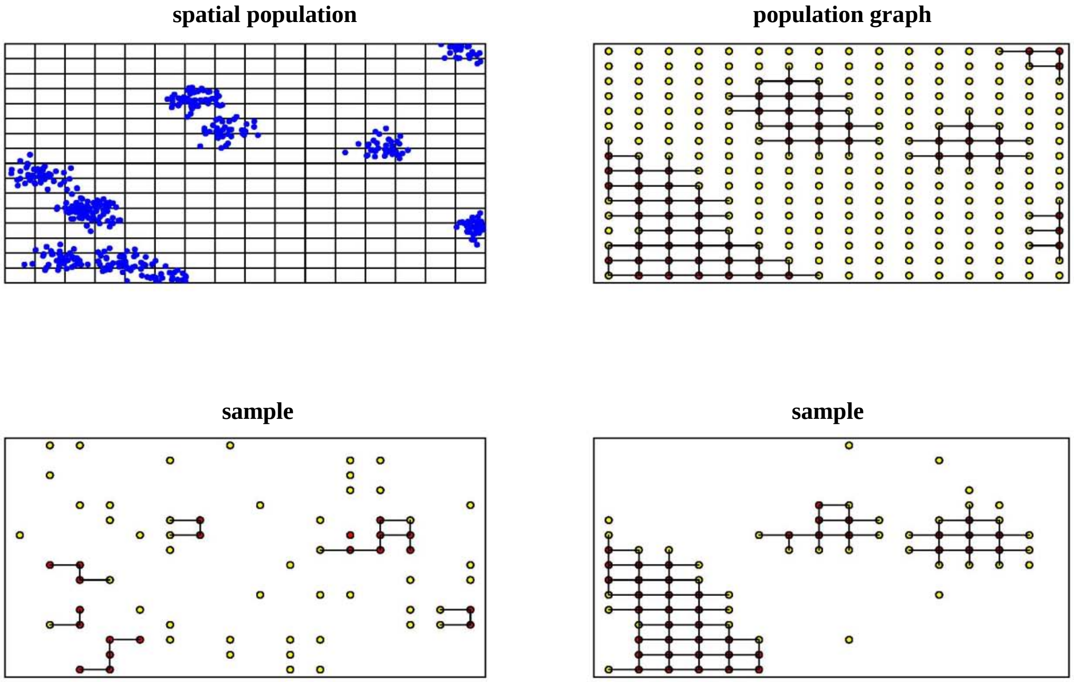


Figure 17 Adaptive web sampling design variations

4. Discussion

Adaptive sampling designs expand considerably the possibilities for sampling strategies. They appear to be especially useful for populations which are otherwise difficult to sample. Network sampling designs are inherently adaptive in most cases and can provide more effective ways to sample populations with network or spatial structure. In this paper the emphasis has been on designs obtaining low mean square error or providing practical means of reaching a hidden population. In other cases the primary objective might be simply to obtain a higher yield sample, that is, a sample having a high total value of the variable of interest. For instance environmental hot spots is where remediation must be made, high risk components of a epidemic related network where treatment or intervention might have the greatest effect. The advantages of an adaptive approach are even more straightforward when the objective is high sample yield.

Fully optimal sampling strategies are in most cases not practical to implement, because of computational complexity and model dependency. A more practical approach is to make improvements over conventional designs with simple adaptive procedures that capture much of the essence, and the choice of design often having much more effect that one inference method versus another.

Simulation analyses with adaptive strategies of different types have tended to lend support to the idea that it is good to have a strong underlying conventional component. Many of the practical strategies have the form of an initial conventional sample with adaptive sampling extending the sample from there through either network or spatial relationships and depending on observed values. Strategies with that type of balance between conventional and adaptive components have in simulations generally performed better than, say, selecting a single unit conventionally and adaptively adding the whole rest of the sample from there. In the simulations most efficient strategies tended to have an initial sampling making up about 60-80 percent of the total sample size. The modest amount of adaptive sampling after that then produced large gains in efficiency. This empirical experience goes along with the characteristic of optimal adaptive strategies, in which there seems to be a push and pull between spreading units far apart or filling in unobserved parts of the study region, corresponding to the conventional component of the simplified designs, and placing new units in the most promising areas, corresponding to the adaptive component in the simplified designs.

Acknowledgements

The author would like to thank the Census Bureau and in particular Patrick Flanagan for hosting a lecture series on which this paper is in part based. Research on the sampling designs described has been supported by the Natural Science and Engineering Research Council, the National Science Foundation, the National Center for Health Statistics, Centers for Disease Control and Prevention, Los Alamos National Laboratories, and the National Institutes of Health.

References

Basu, D. (1969). Role of the sufficiency and likelihood principles in sample survey theory. *Sankhyā*, A, 31, 441-454.

Birnbaum, Z.W., and Sirken, M.G. (1965). Design of sample surveys to estimate the prevalence of rare diseases: Three unbiased estimates. *Vital and Health Statistics*, Series 2, No. 11. Washington: Government Printing Office.

Brin, S., and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Proceedings of the 7th International World Wide Web Conference.* Elsevier, 107-117.

Casella, G., and Robert, C.P. (1996). Rao-Blackwellization of sampling schemes. *Biometrika*, 83.

Chao, C.-T., and Thompson, S.K. (2001). Optimal adaptive selection of sampling sites. *Environmetrics*, 12, 517-538.

Chow, M., and Thompson, S.K. (2003). Estimation with link-tracing sampling designs - A Bayesian approach. *Survey Methodology*, 20, 197-205.

Felix-Medina, M.H., and Thompson, S.K. (2004). Combining link-tracing sampling and cluster sampling to estimate the size of hidden populations. *Journal of Official Statistics*, 20, 19-38.

Frank, O. (1971). *Statistical Inference in Graphs*. Stockholm: Försvarets Forskningsanstalt.

Frank, O. (1977a). Survey sampling in graphs. *Journal of Statistical Planning and Inference*, 1, 235-264.

Frank, O. (1977b). Estimation of graph totals. *Scandinavian Journal of Statistics*, 4, 81-89.

Frank, O. (1978a). Estimating the number of connected components in a graph by using a sampled subgraph. *Scandinavian Journal of Statistics*, 5, 177-188.

Frank, O. (1978b). Sampling and estimation in large social networks. *Social Networks*, 1, 91-101.

Frank, O. (1979). Estimation of population totals by use of snowball samples. In *Perspectives on Social Network Research*, (Eds., P.W. Holland and S. Leinhardt). New York: Academic Press, 319-347.

Frank, O., and Snijders, T. (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics*, 10, 53-67.

Hastings, W.K. (1970). Monte-Carlo sampling methods using Markov chains and their application. *Biometrika*, 57, 97-109.

Heckathorn, D.D. (1997). Respondent driven sampling: A new approach to the study of hidden populations. *Social Problems*, 44, 174-199.

Heckathorn, D.D. (2002). Respondent driven sampling II: Deriving valid population estimates from chain-referral samples of hidden populations. *Social Problems*, 49, 11-34.

Klovdahl, A.S. (1989). Urban social networks: Some methodological problems and possibilities. In *The Small World* (Ed., M. Kochen). Norwood, NJ: Ablex Publishing, 176-210.

Kwanisai, M. (2005). Estimation in Link-Tracing Designs with Subsampling. Ph.D. thesis. The Pennsylvania State University, University Park, PA, U.S.A.

Kwanisai, M. (2006). Estimation in networked populations. *Proceedings of the Survey Research Section*, American Statistical Association, Washington, DC., 3285-3291.

Lovász, L. (1993). Random walks on graphs: A survey. In *Combinatorics*, *Paul Erdös is Eighty* (Eds., D. Miklós, D. Sós and T. Szöni). János Bolyai Mathematical Society, Keszthely, Hungary, 2, 1-46.

Salganik, M.J., and Heckathorn, D.D. (2004). Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological Methodology*, 34, 193-239.

Sirken, M.G. (1970). Household surveys with multiplicity. *Journal of the American Statistical Association*, 63, 257-266.

Sirken, M.G. (1972a). Stratified sample surveys with multiplicity. *Journal of the American Statistical Association*, 67, 224-227.

Sirken, M.G. (1972b). Variance components of multiplicity estimators. *Biometrics*, 28, 869-873.

Thompson, S.K. (1990). Adaptive cluster sampling. *Journal of the American Statistical Association*, 85, 1050-1059.

Thompson, S.K. (2002). *Sampling*, *Second Edition.* New York: John Wiley & Sons, Inc.

Thompson, S.K. (2006a). Targeted random walk designs. *Survey Methodology*, 32, 11-24.

Thompson, S.K. (2006b). Adaptive web sampling. *Biometrics*, 62, 1224-1234.

Thompson, S., and Frank, O. (2000). Model-based estimation with link-tracing sampling designs. *Survey Methodology*, 26, 87-98.

Thompson, S.K., and Seber, G.A.F. (1996). *Adaptive Sampling.* New York: John Wiley & Sons, Inc.

Zacks, S. (1969). Bayes sequential designs of fixed size samples from finite populations. *Journal of the American Statistical Association*, 64, 1342-1349.

Survey Methodology, December 2011
Vol. 37, No. 2, pp. 197-213
Statistics Canada, Catalogue No. 12-001-X

Alternative survey sample designs: Sampling with multiple overlapping frames

Sharon L. Lohr [1]

Abstract

Designs and estimators for the single frame surveys currently used by U.S. government agencies were developed in response to practical problems. Federal household surveys now face challenges of decreasing response rates and frame coverage, higher data collection costs, and increasing demand for small area statistics. Multiple frame surveys, in which independent samples are drawn from separate frames, can be used to help meet some of these challenges. Examples include combining a list frame with an area frame or using two frames to sample landline telephone households and cellular telephone households. We review point estimators and weight adjustments that can be used to analyze multiple frame surveys with standard survey software, and summarize construction of replicate weights for variance estimation. Because of their increased complexity, multiple frame surveys face some challenges not found in single frame surveys. We investigate misclassification bias in multiple frame surveys, and propose a method for correcting for this bias when misclassification probabilities are known. Finally, we discuss research that is needed on nonsampling errors with multiple frame surveys.

Key Words: Bias correction; Dual frame survey; Misclassification; Mode effects; Sampling for rare events; Sampling weights; Small area estimation.

1. Uses of multiple frame surveys

In classical design-based sampling theory, a probability sample is taken from the (single) sampling frame, and the inclusion probabilities in the sampling design can be used to make inferences about the population. Let y_i be a measurement on unit i in the population of N units, let $\mathcal{S}$ denote the set of units in the sample, and let $\pi_i = P$ (unit i is included in the sample). Then the Horvitz-Thompson (1952) estimator of the population total $Y = \sum_{i=1}^{N} y_i$ is $\hat{Y} = \sum_{i \in \mathcal{S}} w_i y_i$, where $w_i = 1 / \pi_i$ is the sampling weight. If the sampling frame includes everyone in the target population, all sampled units respond, and there is no measurement error, then the Horvitz-Thompson estimator is unbiased for Y.

The practical challenges of sampling in the 1940s and 1950s drove the methodological developments of stratified multistage surveys and estimators such as the Horvitz-Thompson estimator. In-person surveys relied on unequal probability sampling to balance interviewer workloads and reduce variances. Response rates were high in many government surveys so that the assumptions for the Horvitz-Thompson estimator were reasonable. We now face new challenges in household surveys. Nonresponse rates are increasing, which means that survey estimates rely more on models. The ethnic and language diversity of a population can result in undercoverage and measurement error. Increasing technological diversity means that different residents may be best reached by different sampling modes; one must then be confident that different sampling modes measure the same quantities. Costs of collecting data have risen greatly, in part due to increasing nonresponse; at the same time, governmental and research demands for data have also risen greatly.

Multiple frame surveys can achieve better population coverage at lower cost. They can be used as part of a structure of modular survey design that relies on different sampling frames to help reduce costs and achieve better coverage. They can also use administrative data efficiently. In this paper, we describe different types of multiple frame surveys and discuss some of the research that is completed and research that may be needed for their use.

One of the earliest multiple frame surveys (aside from early capture-recapture methods) was performed by the Census Bureau in 1949 (Hansen, Hurwitz and Madow 1953). In the Sample Survey of Retail Stores, a probability sample of primary sampling units (psus) was chosen. Within each psu, a list of large retail firms was constructed from records of the Old Age and Survivors Insurance Bureau. All firms on the list were sampled, and an area sample of firms in the psu that were not on the list was taken. In this case, a *screening* dual frame design was employed within each selected psu; units in the list frame were screened out of the area frame before sampling. Thus, the estimator of total sales summed the two estimators within each psu. No new statistical methods were required to estimate total sales in this survey, since essentially a stratified sample was taken in each psu: the firms on the list in the psu formed one stratum, and the firms in the area frame but not on the list in the psu formed the second stratum. The survey resulted in cost savings because it was relatively inexpensive to sample from the firms on the list, yet full coverage was obtained by also using the area frame.

[1] Sharon L. Lohr, School of Mathematical and Statistical Sciences, Arizona State University, Tempe AZ 85287-1804. E-mail: sharon.lohr@asu.edu.

Many agricultural surveys also have used a screening dual frame survey design (González-Villalobos and Wallace 1996). In such a design, farms belonging to the list frame are removed from the area frame before sampling commences. Considerable cost savings can be realized since often the list frame is much less expensive to sample and it contains the largest farms.

In many cases, however, it may not be possible or practical to remove list-frame units from the area frame before sampling. Instead, in an overlapping dual frame survey, independent probability samples are taken from frame A (the area frame) and frame B (the list frame); this is depicted in Figure 1. Rare populations can often be sampled more efficiently using a multiple frame sample (Kalton and Anderson 1986). In an epidemiology study, for example, frame A might be that used for a general population health survey, while frame B might be a list frame of clinics specializing in a certain disease. The sample from frame B is expected to yield a high percentage of persons with the disease of interest, so that sampling will be efficient; the sample from frame A, though more expensive, leads to complete coverage of the population.

In other situations, all frames are incomplete, as considered by Hartley (1962); for example, frame A in Figure 2 might be a frame of landline telephones and frame B might consist of cellular telephone numbers. There are three domains: domain a consists of units in frame A but not in frame B, domain b consists of units in frame B but not in frame A, and domain ab consists of units in both frames. In the telephone context, domain a contains individuals belonging to a landline-only household, domain b consists of individuals who have only a cellular telephone, and domain ab consists of individuals who have both cellular and landline telephones. It is unknown in advance whether a household member sampled using one frame also belongs to the other frame (Brick, Dipko, Presser, Tucker and Yuan 2006); typically, respondents are asked about their cellular and landline telephone usage to determine domain membership.

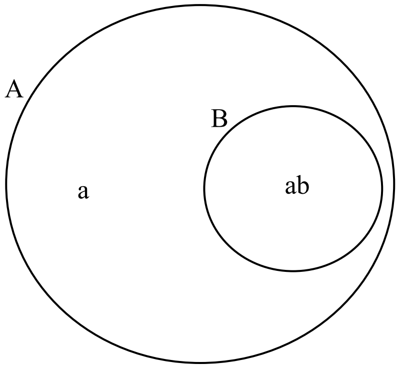


Figure 1 A dual frame design in which frame B is a subset of frame A

More than two frames can be employed as well, as illustrated in Figure 3 for a three-frame survey in which all frames are incomplete. In this situation, there are seven domains. Iachan and Dennis (1993) gave an example of a three-frame survey used to sample the homeless population, where frame A is a list of soup kitchens, frame B is a list of shelters, and frame C consists of street locations. Figure 4 displays a 3-frame survey in which frame A has complete coverage, while overlapping frames B and C are both incomplete but are less expensive to sample. This design has been used for the U.S. Scientists and Engineers Statistical Data System (SESTAT; National Science Foundation 2003) surveys. The same design might be used when A is the frame for a general population survey, B is a landline telephone survey, and C is a cell phone survey.

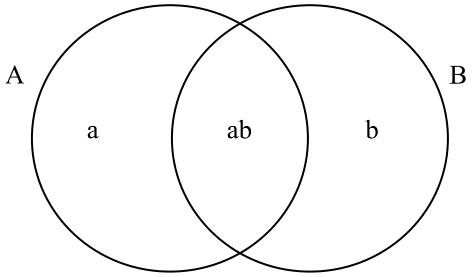


Figure 2 Frames A and B overlap, creating the three domains *a*, *b*, and *ab*

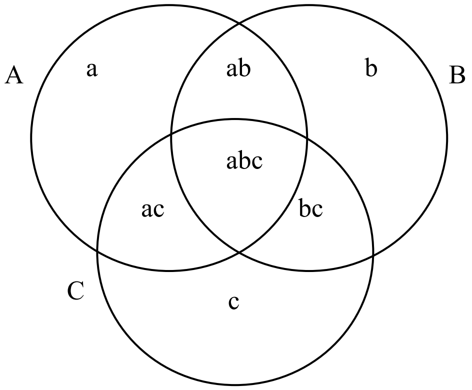


Figure 3 Frames A, B, and C are all incomplete and overlap

There is much potential for using multiple frame designs in household surveys, including:

1. Use of multiple list frames from administrative records.
2. Multiple mode sampling (for example, using independent samples from a cellular telephone frame and a landline telephone frame).

3. Future use of the internet for data collection. Although the internet presents many coverage and domain specification challenges, it is worthy of consideration because of the potential cost savings and ease of data collection and processing.
4. Improved small area estimation. A national survey is supplemented with smaller, localized surveys to obtain higher precision in those areas.
5. Improved estimation for rare populations. A general population survey may be supplemented by a survey from a frame with a high concentration of members of the rare population.
6. Modular survey design. A multiple frame approach can give more flexibility for design of continuing surveys. As particular frames become less expensive to sample, the relative allocation of sample size to the different frames can be modified. The modular approach also allows more flexibility in responding to changing needs for data.

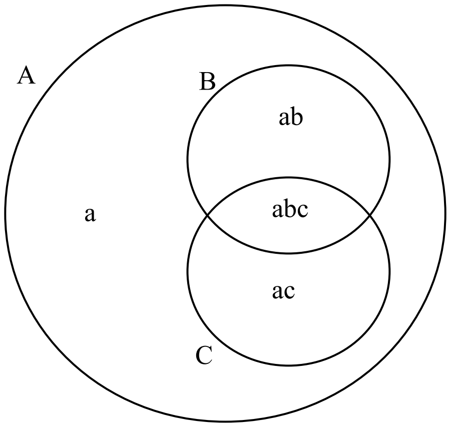


Figure 4 Frame A contains the entire population; frames B and C overlap and are both contained in frame A

The increased flexibility of multiple frame surveys comes at the cost of additional complexity, however. Information from the surveys must be combined to estimate population quantities, and there are many options for estimators. Section 2 summarizes estimators that have been developed for population totals and describes how these modify the sampling weights; Sections 3 and 4 discuss weight calibration and describe how to use survey software packages with multiple frame survey data. Nonsampling errors need to be considered in each frame singly, and in terms of their effect on estimates calculated from the combined information. Section 5 discusses effects of nonresponse and mode effects in multiple frame surveys.

In addition to the nonresponse, undercoverage, and measurement error problems that plague single frame surveys, multiple frame surveys may have domain misclassification. The weight modifications for the estimators in Section 2 depend on the domain membership of the observations. If some observations in domain a are likely to be mistakenly recorded as belonging to domain ab, estimators may have substantial bias. We study effects of domain misclassifycation in Section 6, and propose a new method for adjusting for misclassification bias when misclassification probabilities are known. Finally, Section 7 discusses design issues and Section 8 discusses the potential and challenges of multiple frame surveys.

2. Estimators in overlapping multiple frame surveys

In this section we review estimators for the population total Y from overlapping multiple frame surveys, along with the weight modifications induced by these estimators. For simplicity of notation, we concentrate on dual frame surveys in Section 2.1, and outline extensions to multiple frame surveys in Section 2.2. In a dual frame survey, we can write

$$Y = Y_a + Y_{ab} + Y_b,$$

where Y_a is the total of the population units in domain a, Y_{ab} is the total of the population units in domain ab, and Y_b is the total of the population units in domain b. A special case is estimating the population size $N = N_a + N_{ab} + N_b$, as discussed in Haines and Pollock (1998). We discuss estimating population quantities other than totals and means, and using data from multiple frame surveys in other analyses, in Section 4.

We first set out some desirable properties for estimators from multiple frame surveys.

1. An estimator should be approximately unbiased for the corresponding finite population quantity.
2. Estimators should be internally consistent: that is, if $\hat{Y}_1$ estimates the number of female engineers in the population, $\hat{Y}_2$ estimates the number of male engineers in the population, and $\hat{Y}_3$ estimates the total number of engineers in the population, then we should have $\hat{Y}_1 + \hat{Y}_2 = \hat{Y}_3$. Internal consistency preserves multivariate relationships in the data. In practical terms, internal consistency requires that one set of weights be used for all estimates.
3. An estimator should be efficient, with low mean squared error.
4. An estimator should be of a form that can be calculated with standard survey software such as SUDAAN or SAS PROC SURVEYMEANS. This allows analysts to work with the data without having to write and test new software. In practical

terms, one data file is created from the multiple frame survey. The file includes one column of weights to be used for calculating point estimates, and it contains either variables describing the survey designs for formula-based variance estimation, or columns of replicate weights for replication-based variance estimation.

5. An estimator should, if possible, be robust to nonsampling errors that might occur with multiple frame surveys.

2.1 Estimators and weight adjustments for dual frame surveys

Consider the overlapping dual frame survey depicted in Figure 2, where domain ab is nonempty. A probability sample $\mathcal{S}(A)$ of size n_A is drawn from the N_A units in frame A, and an independent probability sample $\mathcal{S}(B)$ of size n_B is drawn from the N_B units in frame B. Unit i in sample $\mathcal{S}(A)$ has probability of inclusion π_i^A and weight w_i^A, and unit i in sample $\mathcal{S}(B)$ has probability of inclusion π_i^B and weight w_i^B. The weights may be the inverses of the inclusion probabilities, or they may be poststratified to agree with population counts; it is assumed that estimators of population totals are approximately unbiased.

Then $E[\sum_{i\in\mathcal{S}(A)} w_i^A y_i] \approx Y_a + Y_{ab}$ and $E[\sum_{i\in\mathcal{S}(B)} w_i^B y_i] \approx Y_b + Y_{ab}$. Consequently, an estimator that combines the observations from both surveys with the original weights, $\sum_{i\in\mathcal{S}(A)} w_i^A y_i + \sum_{i\in\mathcal{S}(B)} w_i^B y_i$, is biased for the population total Y. If the domain means differ, the corresponding estimator of the population mean may also be biased.

The various estimators for the population total Y that have been proposed in the literature modify the weights so that the estimators are approximately unbiased. The modified weights, shown below for the different estimators, are of the form $\tilde{w}_i^A = m_i^A w_i^A$ and $\tilde{w}_i^B = m_i^B w_i^B$. The population total is then estimated by

$$\hat{Y} = \sum_{i\in\mathcal{S}(A)} \tilde{w}_i^A y_i + \sum_{i\in\mathcal{S}(B)} \tilde{w}_i^B y_i \tag{1}$$

and the population mean $\bar{Y}$ is estimated by $\hat{\bar{Y}} = \hat{Y} / \hat{N}$ where

$$\hat{N} = \sum_{i\in\mathcal{S}(A)} \tilde{w}_i^A + \sum_{i\in\mathcal{S}(B)} \tilde{w}_i^B.$$

The estimators will be approximately unbiased, then, if $m_i^A \approx 1$ for $i \in a$, $m_i^B \approx 1$ for $i \in b$, and $m_i^A + m_i^B \approx 1$ for $i \in ab$. All of the estimators reviewed in this section satisfy the criteria needed for approximate unbiasedness in the absence of nonsampling errors (see Lohr 2009).

Fixed weight adjustments. The simplest weight modification to preserve approximate unbiasedness, described by Hartley (1962), takes

$$m_{i,\theta}^A = \begin{cases} 1 & \text{if } i \in a \\ \theta & \text{if } i \in ab, \end{cases} \qquad m_{i,\theta}^B = \begin{cases} 1 & \text{if } i \in b \\ 1-\theta & \text{if } i \in ab, \end{cases} \tag{2}$$

where $\theta \in [0, 1]$. Using the modified weights $\tilde{w}_i^A = m_{i,\theta}^A w_i^A$ and $\tilde{w}_i^B = m_{i,\theta}^B w_i^B$ in (1), the resulting estimator $\hat{Y}(\theta)$ can also be expressed using the estimated domain totals $\hat{Y}_a^A = \sum_{i\in\mathcal{S}(A), i\in a} w_i^A y_i$, $\hat{Y}_{ab}^A = \sum_{i\in\mathcal{S}(A), i\in ab} w_i^A y_i$, $\hat{Y}_{ab}^B = \sum_{i\in\mathcal{S}(B), i\in ab} w_i^B y_i$, and $\hat{Y}_b^B = \sum_{i\in\mathcal{S}(B), i\in b} w_i^B y_i$. The estimator

$$\begin{aligned} \hat{Y}(\theta) &= \sum_{i\in\mathcal{S}(A)} m_{i,\theta}^A w_i^A y_i + \sum_{i\in\mathcal{S}(B)} m_{i,\theta}^B w_i^B y_i \\ &= \hat{Y}_a^A + \theta \hat{Y}_{ab}^A + (1-\theta) \hat{Y}_{ab}^B + \hat{Y}_b^B \end{aligned} \tag{3}$$

thus estimates the domain total Y_{ab} by a weighted average of the frame A estimator, $\hat{Y}_{ab}^A$, and the frame B estimator, $\hat{Y}_{ab}^B$.

For a fixed value of θ, the estimator $\hat{Y}(\theta)$ gives internal consistency since the same set of adjusted weights is used for all variables. The estimator is simple to use and implement. The efficiency of the estimator depends on the value chosen for θ. Brick *et al.* (2006) used $\theta = 1/2$ in their study of a dual frame survey in which frame A was a landline telephone frame and frame B was a cellular telephone frame, and the value of $\theta = 1/2$ is frequently recommended (see, for example, Mecatti 2007). When $\theta = 0$ or 1, the data in the overlap domain from one of the samples are discarded and the survey becomes a screening dual frame survey.

Optimal estimators. Hartley (1962, 1974) proposed choosing θ in (3) so that the variance of $\hat{Y}(\theta)$ would be minimized. The optimizing value of θ is

$$\theta_H = \frac{V(\hat{Y}_{ab}^B) + \operatorname{Cov}(\hat{Y}_b^B, \hat{Y}_{ab}^B) - \operatorname{Cov}(\hat{Y}_a^A, \hat{Y}_{ab}^A)}{V(\hat{Y}_{ab}^A) + V(\hat{Y}_{ab}^B)}.$$

Since the variances and covariances are generally unknown, they must be estimated from the data, giving

$$\hat{\theta}_H = \frac{\hat{V}(\hat{Y}_{ab}^B) + \widehat{\operatorname{Cov}}(\hat{Y}_b^B, \hat{Y}_{ab}^B) - \widehat{\operatorname{Cov}}(\hat{Y}_a^A, \hat{Y}_{ab}^A)}{\hat{V}(\hat{Y}_{ab}^A) + \hat{V}(\hat{Y}_{ab}^B)}.$$

Skinner and Rao (1996) showed that Hartley's estimator can be calculated using adjusted weights. The weight modifications for Hartley's estimator $\hat{Y}(\hat{\theta}_H)$ are given by (2), substituting $\hat{\theta}_H$ for θ. Since $\hat{\theta}_H$ is consistent for θ_H, Hartley's estimator is asymptotically optimal among all estimators of the form $\hat{Y}_a^A + \hat{Y}_b^B + \theta\hat{Y}_{ab}^A + (1-\theta)\hat{Y}_{ab}^B$. The modified weights $\tilde{w}_{i,H}^A$ and $\tilde{w}_{i,H}^B$ are functions of the variances and covariances of estimated domain totals, however. This has two consequences: (1) the modified weights are random variables, and their variability needs to be accounted for in standard errors of estimators, and (2) the optimal weight modifications will differ for different response variables, leading to internal inconsistency.

Fuller and Burmeister (1972) proposed modifying Hartley's estimator by using additional information about N_{ab}, giving

$$\hat{Y}_{\text{FB}}(\beta) = \hat{Y}_a^A + \hat{Y}_b^B + \beta_1\hat{Y}_{ab}^A + (1-\beta_1)\hat{Y}_{ab}^B + \beta_2(\hat{N}_{ab}^A - \hat{N}_{ab}^B).$$

As with Hartley's estimator, the optimal values $\beta_{1\text{opt}}$ and $\beta_{2\text{opt}}$ are chosen to minimize the variance of $\hat{Y}_{\text{FB}}(\boldsymbol{\beta})$, and are thus functions of the covariances of the domain totals. Substituting consistent estimators $\hat{\beta}_{1\text{opt}}$ and $\hat{\beta}_{2\text{opt}}$ gives the weight adjustments for w_i^A and w_i^B. Lohr and Rao (2000) showed that the Fuller-Burmeister estimator $\hat{Y}_{\text{FB}}$ has the smallest asymptotic variance among the estimators considered. As with the Hartley estimator, however, the modified weights are random variables that differ for different responses, and in complex sampling designs the Fuller-Burmeister estimator is also internally inconsistent.

Pseudo-maximum likelihood (*PML*) *estimators*. To achieve internal consistency Skinner and Rao (1996) proposed a pseudo-maximum likelihood (PML) estimator that uses the same weights for all variables. When N_{ab} is unknown, it is estimated by $\hat{N}_{ab}^{\text{PML}}(\theta)$, which is the smaller of the roots of the quadratic equation

$$\left[\frac{\theta}{N_B} + \frac{1-\theta}{N_A}\right]x^2 - \left[1 + \theta\frac{\hat{N}_{ab}^A}{N_B} + (1-\theta)\frac{\hat{N}_{ab}^B}{N_A}\right]x + \theta\hat{N}_{ab}^A + (1-\theta)\hat{N}_{ab}^B = 0.$$

Skinner and Rao (1996) suggested using the value θ_P for θ that minimizes the asymptotic variance of $\hat{N}_{ab}^{\text{PML}}(\theta)$:

$$\theta_P = \frac{N_a N_B V(\hat{N}_{ab}^B)}{N_a N_B V(\hat{N}_{ab}^B) + N_b N_A V(\hat{N}_{ab}^A)}. \tag{4}$$

Substituting an estimator $\hat{\theta}_P$ for θ_P, the weight adjustments are:

$$m_{i,P}^A = \begin{cases} \dfrac{N_A - \hat{N}_{ab}^{\text{PML}}(\hat{\theta}_P)}{\hat{N}_a^A} & \text{if } i \in a \\ \dfrac{\hat{N}_{ab}^{\text{PML}}(\hat{\theta}_P)}{\hat{\theta}_P\hat{N}_{ab}^A + (1-\hat{\theta}_P)\hat{N}_{ab}^B}\hat{\theta}_P & \text{if } i \in ab, \end{cases}$$

$$m_{i,P}^B = \begin{cases} \dfrac{N_B - \hat{N}_{ab}^{\text{PML}}(\hat{\theta}_P)}{\hat{N}_b^B} & \text{if } i \in b \\ \dfrac{\hat{N}_{ab}^{\text{PML}}(\hat{\theta}_P)}{\hat{\theta}_P\hat{N}_{ab}^A + (1-\hat{\theta}_P)\hat{N}_{ab}^B}(1-\hat{\theta}_P) & \text{if } i \in ab. \end{cases}$$

If the value of θ_P cannot be estimated, for example if the two sampling frames coincide or the design in Figure 1 is used, then one can use an average design effect from each survey in the adjustment, as described in Lohr and Rao (2006). The PML estimator is internally consistent; while not guaranteed to give the smallest mean squared error, it has high efficiency in many survey situations.

Single frame estimators. Bankier (1986) and Kalton and Anderson (1986) proposed estimators of the form in (1) that treat all the observations as though they had been sampled from one frame, with adjusted weights in the intersection domain relying on the inclusion probabilities for each frame. The weight adjustments for the Kalton and Anderson (1986) single frame estimator are:

$$m_{i,S}^A = \begin{cases} 1 & \text{if } i \in a \\ w_i^B / (w_i^A + w_i^B) & \text{if } i \in ab, \end{cases}$$

$$m_{i,S}^B = \begin{cases} 1 & \text{if } i \in b \\ w_i^A / (w_i^A + w_i^B) & \text{if } i \in ab. \end{cases}$$

If $w_i^A = 1/\pi_i^A$ and $w_i^B = 1/\pi_i^B$, the single frame estimator uses $\tilde{w}_{i,S}^A = \tilde{w}_{i,S}^B = 1/(\pi_i^A + \pi_i^B)$ for units in ab. The weight adjustment in domain ab relies on both π_i^A and π_i^B. Thus if a disproportionate stratified random sample is taken from frame B, one must know the frame-B stratum membership for units sampled in $\mathcal{S}(A)$. The adjusted weights from the single frame estimator can be interpreted in terms of inclusion probabilities for sampled units. If the sampling fractions are small, $\tilde{w}_{i,S}^A$ is approximately $1/P$(unit i is included in one of the samples). If each of $\mathcal{S}(A)$ and $\mathcal{S}(B)$ is self-weighting, then the single frame estimator reduces to (3).

The single frame weight modifications are the same for all response variables, so estimators are internally consistent. For complex surveys, however, single frame estimators may not be as efficient as the optimal or PML estimators. Their performance may be improved by raking toward the frame population totals (Skinner 1991).

Pseudo-empirical likelihood (*PEL*) *estimators*. Rao and Wu (2010) proposed empirical likelihood estimators for dual frame surveys. Using $\theta = \theta_P$, the empirical log likelihood function is defined by

$$\begin{aligned} \ell(\mathbf{p}_a, \mathbf{p}_{ab}^A, \mathbf{p}_{ab}^B, \mathbf{p}_b) = {} & \frac{n_A + n_B}{N}\Bigg[\sum_{i\in\mathcal{S}(A),\, i\in a}\frac{N_a}{\hat{N}_a} w_i^A \log(p_{ai}) \\ & + \sum_{i\in\mathcal{S}(A),\, i\in ab}\frac{\theta_P N_{ab}}{\hat{N}_{ab}^A} w_i^A \log(p_{abi}^A) \\ & + \sum_{i\in\mathcal{S}(B), i\in b}\frac{N_b}{\hat{N}_b} w_i^B \log(p_{bi}) \\ & + \sum_{i\in\mathcal{S}(B), i\in ab}\frac{(1-\theta_P)N_{ab}}{\hat{N}_{ab}^B} w_i^B \log(p_{abi}^B)\Bigg], \end{aligned}$$

where θ_P is given in (4). An estimator $\hat{\theta}_P$ is substituted if θ_P is unknown. Then $\ell(\mathbf{p}_a, \mathbf{p}^A_{ab}, \mathbf{p}^B_{ab}, \mathbf{p}_b)$ is maximized subject to

$$\sum\nolimits_{i\in\mathcal{S}(A),i\in a} p_{ai} = 1, \; \sum\nolimits_{i\in\mathcal{S}(A),i\in ab} p^A_{abi} = 1,$$

$$\sum\nolimits_{i\in\mathcal{S}(B),i\in b} p_{bi} = 1, \; \sum\nolimits_{i\in\mathcal{S}(B),i\in ab} p^B_{abi} = 1,$$

and

$$\sum_{i\in\mathcal{S}(A),i\in ab} p^A_{abi} y_i = \sum_{i\in\mathcal{S}(B),i\in ab} p^B_{abi} y_i. \tag{5}$$

When N_{ab} is unknown, the PEL weight modifications are

$$m^A_{i,\text{PEL}} = \begin{cases} \dfrac{p^A_{ai}}{w^A_i}\left\{N_A - \hat{N}^{\text{PML}}_{ab}(\hat{\theta}_P)\right\} & \text{if } i \in a \\ \hat{\theta}_P \dfrac{p^A_{abi}}{w^A_i} \hat{N}^{\text{PML}}_{ab}(\hat{\theta}_P) & \text{if } i \in ab, \end{cases}$$

$$m^B_{i,\text{PEL}} = \begin{cases} \dfrac{p^B_{bi}}{w^B_i}\left\{N_B - \hat{N}^{\text{PML}}_{ab}(\hat{\theta}_P)\right\} & \text{if } i \in b \\ (1-\hat{\theta}_P) \dfrac{p^B_{abi}}{w^B_i} \hat{N}^{\text{PML}}_{ab}(\hat{\theta}_P) & \text{if } i \in ab. \end{cases}$$

The constraint in (5) changes the weights in the overlap domain so that the estimator of Y_{ab} from $\mathcal{S}(A)$ is forced to equal the estimator of Y_{ab} from $\mathcal{S}(B)$. This constraint, however, results in a different set of weights for each response variable. The PEL estimator thus is not internally consistent. Rao and Wu (2010) presented an alternative multiplicity version in which the weight adjustments do not depend on y; in the absence of auxiliary information, this estimator is the same as $\hat{Y}(1/2)$ in (3).

2.2 Weight adjustments with three or more frames

In the general case, suppose there are Q frames, denoted $A_1, ..., A_Q$. Let $\mathcal{S}(A_q)$ denote the probability sample from frame A_q, for $q = 1, ..., Q$. Unit i in sample $\mathcal{S}(A_q)$ has probability of inclusion $\pi^{A_q}_i$ and weight $w^{A_q}_i$. There are a total of D distinct domains.

A multiple frame estimator generalizing (1) is of the form

$$\hat{Y} = \sum_{q=1}^{Q} \sum_{i\in\mathcal{S}(A_q)} m^{A_q}_i w^{A_q}_i y_i,$$

where $m^{A_q}_i$ is the weight adjustment for observation i in $\mathcal{S}(A_q)$. A fixed weight estimator sets weight adjustments $m^{(A_q,d)}$ for each frame and domain, with the constraints that $m^{(A_q,d)} \geq 0$ ($m^{(A_q,d)}$ is assumed to equal 0 if domain d is not part of frame A_q) and $\sum_{q=1}^{Q} m^{(A_q,d)} = 1$ for $d = 1, ..., D$. Then, $m^{A_q}_i = m^{(A_q,d)}$ when observation i from $\mathcal{S}(A_q)$ is in domain d. A simple choice, which generalizes the fixed weight dual frame estimator $\hat{Y}(1/2)$ in (3), takes $m^{(A_q,d)} = [1/\text{number of frames that contain domain } d]$; this is called the multiplicity estimator by Mecatti (2007). Other choices include setting $m^{(A_q,d)} = 1$ in exactly one frame and 0 for the other frames, resulting in screening estimators.

Many of the properties from the dual frame situation extend to the case of three or more frames; multiple frame versions of the estimators in Section 2.1 were studied by Hartley (1974), Lohr and Rao (2006), and Mecatti (2007). How do the multiple frame estimators satisfy the criteria set out at the beginning of this section? All of the estimators – fixed weight, optimal, PML, PEL, and single frame – are approximately unbiased for population totals when sufficiently large samples are taken in the frames. The fixed weight, PML, and single frame estimators are internally consistent; the optimal Hartley-type and Fuller-Burmeister-type estimators in Lohr and Rao (2006) and a multiple-frame extension of the PEL estimator of Rao and Wu (2010) are not internally consistent. While the optimal estimators are asymptotically efficient, they are often unstable in small or moderate samples with three or more frames because the optimal estimated weight modifications are functions of large estimated covariance matrices. The optimal and PEL estimators are ill suited for use with standard survey software because they require a different set of weights for each response variable.

We recommend that one of the internally consistent estimators – fixed weight, PML, or single frame – be used in practice. Lohr and Rao (2006) concluded that the PML estimator has small mean squared error in many survey circumstances, and thus is a good choice for a survey that is conducted only once. With repeated surveys, though, the simplicity and transparency of a fixed weight estimator may be preferred. Fixed weight adjustments may make year-to-year comparisons easier in an annual survey where the domain proportions are relatively constant over time. Fixed weight estimators are also more amenable to weight adjustments for nonresponse and domain misclassification (see Sections 5.1 and 6.1). If fixed weight adjustments can be chosen that are close to the optimal weight adjustments for important responses, perhaps by using estimated design effects from previous surveys, the fixed weight estimator will have mean squared error close to that of the optimal and PML estimators.

3. Postratification to population controls

All of the estimators in Section 2 modify the original sampling weights. As a result, some properties of the original weights may be lost. For example, if a stratified random sample is taken in frame A, the modified weights will not necessarily have the property that the sum of the weights in a stratum equals the stratum population size.

Bankier (1986), in the original development of single frame estimation methods, suggested raking the single frame weights, $\tilde{w}_{i,S}^{A}$ and $\tilde{w}_{i,S}^{B}$, to stratum totals so that the adjusted weights $\tilde{w}_{i,S}^{A,\text{adj}}$ and $\tilde{w}_{i,S}^{B,\text{adj}}$ satisfy

$$\sum\nolimits_{i \in S_{Ah}} (\tilde{w}_{i,S}^{A,\text{adj}} + \tilde{w}_{i,S}^{B,\text{adj}}) = N_{Ah},$$

where S_{Ah} represents the sampled units from either frame in stratum h of frame A, and N_{Ah} is the population size of that stratum. Bankier (1986) and Skinner (1991) used raking ratio estimation to calibrate single frame estimators to the frame population sizes N_A and N_B. Kott, Amrhein and Hicks (1998) proposed using the least squares calibration methods of Deville and Särndal (1992) for calibrating weights to population totals such as stratum sizes.

For the PML estimator, Lohr and Rao (2000) recommended combining the samples first and then using calibration methods to adjust to population as well as separate-frame population totals. When nonresponse is present and a fixed weight estimator is used, Brick, Cervantes, Lee and Norman (2011) concluded that it is preferable to poststratify the individual samples first, and then combine the samples. In some situations, it is most efficient to poststratify both before and after combining samples; in other situations, poststratification can increase bias (see Section 6). Decisions about poststratification need to be made based on the mean squared error, which includes effects of nonsampling errors, and not just on the sampling variance.

4. Analyzing multiple frame surveys with survey software

4.1 Point estimation with survey software

Only internally consistent weight adjustments are suitable for use with survey software when there are multiple responses of interest. Each of the internally consistent methods presented in Section 2.1 results in one vector of adjusted weights for each sample. These may then be concatenated to form one vector of weights: $\tilde{\mathbf{w}} = [\tilde{w}_i^{A_1}, i \in \mathcal{S}(A_1), ..., \tilde{w}_i^{A_Q}, i \in \mathcal{S}(A_Q)]$. Let $\mathbf{y}$ be the corresponding vector of observations, formed by concatenating the observations from samples $\mathcal{S}(A_1)$ through $\mathcal{S}(A_Q)$. Then $\hat{Y} = \tilde{\mathbf{w}}'\mathbf{y}$. From a user's perspective, once the modified weights are constructed, the procedure followed to find point estimates of population totals and means is the same as in a single frame survey.

The modified weights from an internally consistent procedure can be used to estimate any population quantity. Let $F(y)$ be the cumulative distribution function for the population, with

$$F(y) = \sum_{i=1}^{N} I(y_i \le y) \Big/ N,$$

where $I(y_i \le y) = 1$ if $y_i \le y$ and 0 otherwise. In a single frame survey, $F(y)$ is estimated by the empirical cumulative distribution function

$$\hat{F}(y) = \sum\nolimits_{i \in \mathcal{S}} w_i I(y_i \le y) \Big/ \sum\nolimits_{i \in \mathcal{S}} w_i.$$

The modified weights may be used to estimate $F(y)$ in a multiple frame survey:

$$\hat{F}(y) = \sum_{q=1}^{Q} \sum_{i \in \mathcal{S}(A_q)} \tilde{w}_i^{A_q} I(y_i \le y) \Big/ \sum_{q=1}^{Q} \sum_{i \in \mathcal{S}(A_q)} \tilde{w}_i^{A_q}.$$

The denominator is approximately unbiased for N, and the numerator is approximately unbiased for $\sum_{i=1}^{N} I(y_i \le y)$. Any functional of the cumulative distribution function may then be estimated using $\hat{F}(y)$: the mean, $\int y dF(y)$, the median m satisfying $F(m) \approx 1/2$, or any other quantity.

Since the estimators with modified weights are approximately unbiased for population means and totals, they are also approximately unbiased for smooth functions of population means such as ratios and regression coefficients. Any population quantity that could be estimated using the weights from a single frame survey can be estimated analogously using the adjusted weight vector for the multiple frame survey.

4.2 Variance estimation with survey software

Knowledge of the survey designs is needed to calculate standard errors. Variance estimation is straightforward for the estimator in (3), where the weight adjustments do not depend on the data. In that situation,

$$V[\hat{Y}(\theta)] = V\left[\sum_{i \in \mathcal{S}(A)} \tilde{w}_i^{A} y_i\right] + V\left[\sum_{i \in \mathcal{S}(B)} \tilde{w}_i^{B} y_i\right],$$

where $\tilde{w}_i^{A}$ and $\tilde{w}_i^{B}$ are defined below (2). Create the data set by concatenating the observations from $\mathcal{S}(A)$ and $\mathcal{S}(B)$ as in Section 4.1, using $\tilde{w}_i^{A}$ and $\tilde{w}_i^{B}$ as the weights. Define the stratification variable for the combined sample as the combination of categories given by the frame indicator variable, the frame-A stratification variable, and the frame-B stratification variable. Define the first-stage clustering variable for the combined sample similarly as the combination of categories of the individual frame clustering variables. Then, standard survey software may be used to estimate population means and totals using the modified weights, and to estimate variances using the stratification and clustering variables from the combined samples.

Variance estimation is more complicated when the weight modifications m_i^{A} or m_i^{B} depend on quantities that are estimated from the sample, as in the PML estimator, or when the combined sample is poststratified or calibrated to population quantities. Linearization, jackknife, and bootstrap methods may then be used to estimate variances.

In the following, we summarize methods that can be used for variance estimation if the psus from the frames are selected independently. If samples from the different frames share psus, other methods must be used. If, for example, psus are selected from the population, and a dual frame design is used within each selected psu, point estimators for psu totals can be calculated using one of the methods described in Section 2. Then standard replication methods can be used to calculate a with-replacement variance estimator.

Under regularity conditions, the linearization and jackknife methods are consistent for estimating the variance of a population characteristic τ that can be written as $\tau = g(\mathbf{A}, \mathbf{B})$, where $\mathbf{A}$ is a vector of population totals from frame A, $\mathbf{B}$ is a vector of population totals from frame B, and g is a twice continuously differentiable function (Skinner and Rao 1996; Lohr and Rao 2000). The vector $\mathbf{A}$ is estimated from $\mathcal{S}(A)$ by $\hat{\mathbf{A}}$, with estimated covariance matrix $\hat{\Sigma}_A$; similarly, $\hat{\mathbf{B}}$ estimates $\mathbf{B}$ from $\mathcal{S}(B)$, with $\hat{V}(\hat{\mathbf{B}}) = \hat{\Sigma}_B$. The linearization estimator of the variance of $\hat{\tau} = g(\hat{\mathbf{A}}, \hat{\mathbf{B}})$ is

$$\hat{V}_L(\hat{\tau}) = g'_A \hat{\Sigma}_A g_A + g'_B \hat{\Sigma}_B g_B,$$

where g_A is the vector of partial derivatives of $g(\mathbf{A}, \mathbf{B})$ with respect to the components of $\mathbf{A}$ and g_B is the corresponding vector of partial derivatives for frame B. Demnati, Rao, Hidiroglou and Tambay (2007) derived linearization estimators of the variance for multiple frame surveys by taking derivatives of a function of the weights rather than of the means. Linearization methods require that the derivatives be calculated separately for each estimator that is considered, and these calculations can be cumbersome. For that reason, it may be preferred to use replication methods if multiple frame surveys are adopted.

Suppose a stratified multistage sample with H strata is taken from frame A, where stratum h has $\tilde{n}_h^A$ primary sampling units. An independent stratified multistage sample with L strata is taken from frame B, where stratum l has $\tilde{n}_l^B$ primary sampling units. The jackknife estimator of the variance can be calculated by creating a total of $\sum_{h=1}^{H} \tilde{n}_h^A + \sum_{l=1}^{L} \tilde{n}_l^B$ replicate weight columns (Lohr and Rao 2000). The replicate weights for the column corresponding to the deletion of psu i from stratum h in $\mathcal{S}_A$ are formed by:

$$\tilde{w}_{k(hi)}^A = \begin{cases} \dfrac{\tilde{n}_h^A}{\tilde{n}_h^A - 1} \tilde{w}_k^A & \text{if unit } k \text{ is in stratum } h \text{ but not in psu } i, \\ 0 & \text{if unit } k \text{ is in psu } i \text{ of stratum } h, \\ \tilde{w}_k^A & \text{if unit } k \text{ is in stratum } g \neq h. \end{cases}$$

The jackknife coefficient for this column is the multiplier $(\tilde{n}_h^A - 1) / \tilde{n}_h^A$. The column of replicate weights corresponding to the deletion of psu j from stratum l in $\mathcal{S}_B$ is formed similarly, with jackknife coefficient $(\tilde{n}_l^B - 1) / \tilde{n}_l^B$. With more than two frames, additional columns of replicate weights are added corresponding to the deleted psus from those samples. Weights for a bootstrap method of variance estimation (see Lohr 2007) can be defined similarly.

Multiple frame replication variance methods can be used with standard survey packages that allow replicate weights. If desired, each column in the replicate weights can be post-stratified to population and frame totals, so that the post-stratification is accounted for in the variance estimation.

One challenge with replication variance methods is that the number of columns of replicate weights needed may be very large if a simple random sample or stratified random sample is taken in one of the frames. For the bootstrap, we have found that for some surveys at least 500 bootstrap iterations are needed for variance estimates with dual frame surveys, which again may be excessive. It is possible that combined strata variance estimation, as discussed in Lu, Brick and Sitter (2006), may be used with multiple frame surveys to reduce the number of replicates needed.

5. Nonsampling errors

Multiple frame surveys often have better population coverage than a single frame surveys. When all frames are incomplete, as in Figure 3, any one of frames A, B, or C, if used as the sole sampling frame, would have severe undercoverage. The multiple frame survey design ensures that all units in the overlapping frames have a positive probability of inclusion.

Like all surveys, multiple frame surveys are subject to nonsampling errors. They have nonresponse, which may differ in the different frames. While the union of the frames may have better coverage than a single frame, there may still be undercoverage of the target population. Estimators for multiple frame surveys are also sensitive to domain misclassification and biases that might result from different administration methods or modes in the component surveys. We discuss nonresponse and mode effects in this section, and study effects of domain misclassification in Section 6.

5.1 Nonresponse

In any survey, nonresponse can result in biased estimates of population totals and other quantities. Different nonresponse rates in the samples from the two frames can affect the point estimates of the population total given in Section 2; additionally, nonresponse can affect the weight adjustments prescribed by some of the methods.

Kennedy (2007) discussed a problem that has occurred when frame A consists of landline telephone numbers and frame B has cellular telephone numbers: the units in the intersection domain ab who were interviewed by cell phone differed from those in ab who were interviewed on the landline phone. For example, it was estimated that 18% of the intersection units were aged 18-25 in the frame-B sample, while it was estimated that only 8% of the intersection units were aged 18-25 using the frame-A sample. The difference was ascribed to nonresponse: it was thought that persons who predominantly use cellular telephones (and thus are difficult to reach through a landline survey) tend to be younger. Kennedy (2007) suggested raking using estimated relative telephone usage (*i.e.*, whether most of calls are on landline or cellular telephone).

Brick *et al.* (2011) proposed two methods for nonresponse adjustment in dual frame cellular/landline telephone surveys with fixed weight estimators. They considered a setup in which the overlap domain has two groups: households that receive all or nearly all of their calls on cellular telephones (cell-mainly), and the remaining households in the overlap domain (landline-mainly). The first method, which does not require external estimates of control totals, sets the value of θ in the fixed weight adjustment estimator to reduce the nonresponse bias by using the response rates for the cell-mainly and landline-mainly households in each sample. The second method requires poststratification control totals for the cell-mainly and landline-mainly groups in the overlap domain, N_{1ab} and N_{2ab}, and estimates the population total in domain ab by

$$\sum_{g=1}^{2}\left[\theta_g \frac{N_{gab}}{\hat{N}_{gab}^{A}} \hat{Y}_{gab}^{A} + (1-\theta_g) \frac{N_{gab}}{\hat{N}_{gab}^{B}} \hat{Y}_{gab}^{B}\right],$$

where $\hat{Y}_{gab}^{A}$ represents the estimated total of group g in domain ab from $\mathcal{S}(A)$, the other totals are defined similarly, and $0 \le \theta_g \le 1$ for $g = 1, 2$.

5.2 Mode effects

In some cases, multiple frame may also mean multiple mode. De Leeuw (2008) compared the advantages and disadvantages of different sampling modes, and summarized empirical research on mode biases. Persons may give different responses when presented with questions in a visual form than when presented with questions in an auditory form, resulting in mode bias. Mode effects that occur in single frame surveys will also occur in multiple frame surveys. If different modes are used in different frames, it is challenging to separate mode effects from other nonsampling errors.

Many of the multiple frame survey estimators combine estimates from the overlap domains, and these methods assume that the estimators of Y_{ab} from the component surveys both estimate the same quantity. If, however, the frame A survey is conducted in person while the frame B survey is conducted by telephone, it is possible that a census of the domain ab from frame B would give a different domain total than a census from frame A.

One possibility to investigate mode effects is to conduct the frame B survey using a split sample, *e.g.*, partly in person and partly by telephone, but that would reduce the cost savings from the dual frames. Careful pretesting can mitigate the mode effects. Research is needed in this area; the same problem of mode effects, of course, occurs in single frame surveys such as the American Community Survey in which nonresponse follow-up is done by different mode than the original sample (see Citro and Kalton 2007). The methods presented in de Leeuw, Hox and Dillman (2008) for designing surveys for multiple modes also apply in the multiple frame setting.

Vannieuwenhuyze, Loosveldt and Molenberghs (2011) presented a method for distinguishing mode effects from selection effects when a supplemental single-mode survey is available. They noted, however, that the method requires the strong assumption that the coverage and nonresponse errors are equivalent for both surveys. If this assumption is met for a dual frame survey so that the samples in the overlap domain from frames A and B represent the same population, and if domain classification is correct, the mode effect can be estimated from the overlap domain as $D_{ab} = \hat{Y}_{ab}^{A} - \hat{Y}_{ab}^{B}$. A difference that is significantly different from 0 indicates presence of a mode effect if there are no other nonsampling errors. If other nonsampling errors are present, a large value of D_{ab} does not provide information about the cause of the difference; experimentation is needed to distinguish possible causes.

6. Domain misclassification and bias adjustment

The estimators discussed in Section 2 construct weights for the observations based on domain membership. Thus in the estimator $\hat{Y}(\theta)$ in (3), the weight multiplier of an observation from sampling frame A is 1 if the observation is in domain a, and is θ if the observation is in domain ab, in order to account for the multiplicity of sampling.

In practice, domain membership may not be clear. For the situation in Figure 1, it may be unknown whether a respondent in an area frame also belongs to the list frame. If frame A is an area frame and frame B is an internet frame, for example, the only way to determine whether an individual sampled from frame A is also in frame B may be to ask the person about internet access, and the person might not give the correct response.

If matching or record linkage is used to determine frame membership, imperfect matching can also misclassify observations. Lesser and Kalsbeek (1999) discussed nonsampling errors that occur in dual frame surveys that have been conducted by the U.S. National Agricultural Statistics Service. Domain misclassification can occur if a farm sampled in the area frame is incorrectly classified with respect to its list frame membership. In landline/cellular dual frame telephone surveys, it is challenging to determine whether a person in one frame is also in the other frame (Kennedy 2007). A person reached in a landline telephone sample may also have a cell phone, but rarely take calls on the cell phone. While technically in the overlap domain, that person is virtually unreachable in the cell phone survey. Some landline/cellular surveys ask respondents about the relative amounts of cellular or landline telephone usage, but misclassification can occur.

In practice, we expect domain misclassification to be related to responses of interest; we also expect that in many situations, misclassification is more likely to occur in certain directions. In longitudinal dual frame surveys, domain misclassification can have greater effects than in cross-sectional surveys (Lu and Lohr 2010). In some situations, the domain indicator can be missing or unavailable. Clark, Winglee and Liu (2007) investigated logistic regression and record-linkage methods for predicting the domain of an observation with missing domain information.

6.1 Misclassification bias adjustments

If domain misclassification is severe, each method for modifying the survey weights to adjust for multiplicity can result in biased estimates of population quantities. In this section we derive a correction for the domain misclassification bias of the fixed weight estimator of Section 2.2 when misclassification probabilities are known. Let the D-vector $\boldsymbol{\delta}_i^{A_q}$ denote the true domain membership for observation i of frame A_q, containing a 1 in position d if observation i is in domain d, and 0 elsewhere. Let $\mathbf{Y} = (Y_1, \ldots, Y_D)'$ denote the vector of population totals for the D domains. For an overlapping dual frame survey, $\mathbf{Y} = (Y_a, Y_{ab}, Y_b)'$; for a three-frame survey, $\mathbf{Y} = (Y_a, Y_{ab}, Y_{ac}, Y_{abc}, Y_b, Y_{bc}, Y_c)'$. If there is no domain misclassification,

$$\hat{\mathbf{Y}}^{A_q} = \sum\nolimits_{i \in \mathcal{S}(A_q)} \boldsymbol{\delta}_i^{A_q} w_i^{A_q} y_i$$

is the corresponding estimator of $\mathbf{Y}$ from $\mathcal{S}(A_q)$. For fixed weight adjustment vector $\mathbf{m}^{A_q} = (m^{(A_q, 1)}, \ldots, m^{(A_q, D)})'$ in frame A_q, satisfying $\sum_{q=1}^{Q} \mathbf{m}^{(A_q, d)} = 1$, then $E[\sum_{q=1}^{Q} (\mathbf{m}^{A_q})' \hat{\mathbf{Y}}^{A_q}] = Y$.

Now suppose there is misclassification. Let $\boldsymbol{\eta}_i^{A_q}$ denote the observed classification for observation i in $\mathcal{S}$. We can write $\boldsymbol{\eta}_i^{A_q} = (\mathbf{M}_i^{A_q})' \boldsymbol{\delta}_i^{A_q}$, where $\mathbf{M}_i^{A_q}$ is a $D \times D$ matrix containing a 1 in position (d, e) if observation i in true domain d is (mis)classified to domain e, and 0 elsewhere.

To allow for differential misclassification within domains, we posit a structure in which the misclassification probabilities can differ for subpopulations in a frame. In a landline/cellular survey, for example, some age groups may be known to have higher misclassification probabilities than others. Chambers, Chipperfield, Davis and Kovačević (2008) used a similar grouping approach to correct for record linkage errors. Suppose the population can be divided into G groups, $g = 1, \ldots, G$, in which the misclassification probabilities are known for each frame A_q. Let $\phi_g^{A_q}(d, e)$ denote the probability that an observation in group g with true domain d is classified into domain e in sample $\mathcal{S}(A_q)$, and let $\boldsymbol{\Phi}_g^{A_q}$ be the $D \times D$ matrix with entries $\phi_g^{A_q}(d, e)$. For observation i belonging to group g and true domain d, assume that row d of $\mathbf{M}_i^{A_q}$ is generated as a multinomial random variable of size 1 with probabilities in row d of the expected misclassification matrix $\boldsymbol{\Phi}_g^{A_q}$, and that all $\mathbf{M}_i^{A_q}$ are independent of each other and of the sample inclusion variables. We thus have G matrices of misclassification probabilities for frame A_q, $\boldsymbol{\Phi}_1^{A_q}, \ldots, \boldsymbol{\Phi}_G^{A_q}$. Denote the vector of population totals for group g by $\mathbf{Y}(g) = \sum_{i=1}^{N} \boldsymbol{\delta}_i^{A_q} \chi_i(g) y_i$, where $\chi_i(g) = 1$ if observation i is in group g and 0 otherwise.

With the observed domain classifications $\boldsymbol{\eta}_i^{A_q}$, the design-weighted estimator of the vector of domain totals in group g is

$$\begin{aligned}\hat{\mathbf{Y}}^{A_q}(\text{mis}, g) &= \sum_{i \in \mathcal{S}(A_q)} \boldsymbol{\eta}_i^{A_q} \chi_i(g) w_i^{A_q} y_i \\ &= \sum_{i \in \mathcal{S}(A_q)} (\mathbf{M}_i^{A_q})' \boldsymbol{\delta}_i^{A_q} \chi_i(g) w_i^{A_q} y_i,\end{aligned}$$

so that $E[\hat{\mathbf{Y}}^{A_q}(\text{mis}, g)] = (\boldsymbol{\Phi}_g^{A_q})' \mathbf{Y}(g)$.

Now consider a new vector of weight adjustments $\tilde{\mathbf{m}}_g^{A_q} = (\tilde{m}_g^{(A_q, 1)}, \ldots, \tilde{m}_g^{(A_q, D)})'$ for group g in frame A_q. Then

$$E\left[\sum_{g=1}^{G} \sum_{q=1}^{Q} (\tilde{\mathbf{m}}_g^{A_q})' \hat{\mathbf{Y}}^{A_q}(\text{mis}, g)\right] = \sum_{g=1}^{G} \sum_{q=1}^{Q} (\boldsymbol{\Phi}_g^{A_q} \tilde{\mathbf{m}}_g^{A_q})' \mathbf{Y}(g).$$

Since $\sum_{g=1}^{G} \sum_{q=1}^{Q} (\mathbf{m}^{A_q})' \mathbf{Y}(g) = Y$, the bias will be eliminated under this model when

$$\tilde{\mathbf{m}}_g^{A_q} = (\boldsymbol{\Phi}_g^{A_q})^{+} \mathbf{m}^{A_q}, \qquad (6)$$

where $(\boldsymbol{\Phi}_g^{A_q})^{+}$ is the Moore-Penrose inverse of $\boldsymbol{\Phi}_g^{A_q}$, obtained by taking the inverse of the nonzero rows and columns of $\boldsymbol{\Phi}_g^{A_q}$.

Replacing weight adjustments $\mathbf{m}^{A_q}$ by $\tilde{\mathbf{m}}_g^{A_q}$ eliminates the bias under the multinomial misclassification model but inflates the variance. For frame A_q,

$$V\left[\sum_{g=1}^{G} (\tilde{\mathbf{m}}_g^{A_q})' \,\hat{\mathbf{Y}}^{A_q}(\text{mis}, g)\right]$$

$$= E\left[V\left(\sum_{g=1}^{G} \sum_{i\in\mathcal{S}(A_q)} \{(\mathbf{\Phi}_g^{A_q})^{+}\mathbf{m}^{A_q}\}' \,(\mathbf{M}_i^{A_q})' \right.\right.$$

$$\left.\left. \delta_i^{A_q} \chi_i(g)\, w_i^{A_q} y_i \mid \mathcal{S}(A_1), \ldots, \mathcal{S}(A_Q) \right)\right]$$

$$+ V\left[E\left(\sum_{g=1}^{G} \sum_{i\in\mathcal{S}(A_q)} \{(\mathbf{\Phi}_g^{A_q})^{+}\mathbf{m}^{A_q}\}' \,(\mathbf{M}_i^{A_q})' \right.\right.$$

$$\left.\left. \delta_i^{A_q} \chi_i(g)\, w_i^{A_q} y_i \mid \mathcal{S}(A_1), \ldots, \mathcal{S}(A_Q)\right)\right]$$

$$= \sum_{g=1}^{G} [(\mathbf{\Phi}_g^{A_q})^{+}\mathbf{m}^{A_q}]' \, E\left[\sum_{i\in\mathcal{S}(A_q)} \chi_i(g)(w_i^{A_q} y_i)^2 \right.$$

$$\left. \{\text{diag}[(\mathbf{\Phi}_g^{A_q})'\delta_i^{A_q}] - (\mathbf{\Phi}_g^{A_q})'\delta_i^{A_q} (\delta_i^{A_q})' \mathbf{\Phi}_g^{A_q}\} \right] (\mathbf{\Phi}_g^{A_q})^{+}\mathbf{m}^{A_q}$$

$$+ V\left[\sum_{i\in\mathcal{S}(A_q)} \{\mathbf{m}^{A_q}\}' \,\delta_i^{A_q} w_i^{A_q} y_i \right].$$

The second term is the variance of the contribution from frame A_q when the units are classified correctly. The first term is zero only when $\mathbf{\Phi}_g^{A_q}$ is diagonal for all g, *i.e.*, there is no misclassification.

The weight adjustments in (6) may be extended to the case in which the original fixed weights $\mathbf{m}^{A_q}$ vary for the groups, as long as $\sum_{q=1}^{Q} m_g^{(A_q, d)} = 1$ for each domain. Note that the bias correction method in this section is proposed only for the fixed weight estimators, and not for the PML, PEL, or optimal estimators where the multiplicity weights depend on the data. The bias correction depends on the correct specification of the misclassification probabilities. If the misclassification probabilities are estimated from another survey, the operational methods of the surveys must be similar.

6.2 Simulation study

Lohr and Rao (2006) found in simulation studies that the PML estimator has smaller mean squared error than the other estimators when random misclassification is present, but this is due largely to the smaller variance of that estimator. To study sensitivity of estimators to other forms of domain misclassification, we performed a simulation study for two- and three-frame surveys. The population for domain d was generated using the model $y_{ij} = \mu_d + \alpha_i + \varepsilon_{ij}$ for $i = 1, \ldots, N_d$ and $j = 1, \ldots, 5,$ with $\alpha_i \sim N(0, 1)$ and $\varepsilon_{ij} \sim N(0, 1)$ generated independently, and then probability samples were drawn from this population.

For the two-frame study, the domain means are $\mu_a = -1, \mu_{ab} = 0, \mu_b = 2$ and factors in the simulation are:

1. Sample size: 100 or 200 from each frame.
2. Cluster sample or simple random sample drawn from frame A. A cluster sample was drawn by selecting a simple random sample of $n_A/5$ of the groups used in generating the population.
3. Misclassification probabilities for frame A (all probabilities not listed are 0):
   a. $\phi_{aa}^{A} = 1,\ \phi_{ab,ab}^{A} = 1$ (no misclassification);
   b. $\phi_{aa}^{A} = 0.9,\ \phi_{a,ab}^{A} = 0.1,\ \phi_{ab,ab}^{A} = 1;$
   c. $\phi_{aa}^{A} = 0.9,\ \phi_{a,ab}^{A} = 0.1,\ \phi_{ab,ab}^{A} = 0.9, \phi_{ab,a}^{A} = 0.1;$
   d. $\phi_{aa}^{A} = 1,\ \phi_{ab,ab}^{A} = 0.9,\ \phi_{ab,a}^{A} = 0.1.$
4. Misclassification probabilities for frame B:
   a. $\phi_{bb}^{B} = 1, \phi_{ab,ab}^{B} = 1$ (no misclassification);
   b. $\phi_{bb}^{B} = 0.8,\ \phi_{b,ab}^{B} = 0.2,\ \phi_{ab,ab}^{B} = 1;$
   c. $\phi_{bb}^{B} = 0.8,\ \phi_{b,ab}^{B} = 0.2,\ \phi_{ab,ab}^{B} = 0.9, \phi_{ab,b}^{B} = 0.1;$
   d. $\phi_{bb}^{B} = 1,\ \phi_{ab,ab}^{B} = 0.8,\ \phi_{ab,b}^{B} = 0.2.$
5. Population sizes: $N_a = N_b = N_{ab} = 25{,}000;$ $N_a = N_b = 10{,}000,$ $N_{ab} = 55{,}000;$ $N_a = 25{,}000,$ $N_{ab} = 40{,}000,$ $N_b = 10{,}000.$

Ten thousand replicates were run for each combination of the factors, giving the Monte Carlo estimate of bias a standard error of approximately 100. We studied all estimators in Section 2, including $\hat{Y}(1/2), \hat{Y}(2/3),$ and $\hat{Y}(1)$ from (3). We also examined poststratified estimators that could be employed when the domain population counts $N_a, N_{ab},$ and N_b are known: estimators with subscript "post1" apply poststratification to the two samples first and then combine the samples, and estimators with subscript "post2" combine the samples first and then poststratify to the domain population counts. The bias corrected estimators $\hat{Y}(1/2)_{bc}$ and $\hat{Y}(2/3)_{bc}$ modify the initial fixed weights corresponding to $\theta = 1/2$ and $\theta = 2/3$ using (6). With misclassification pattern (b) in frame A, for example, the bias-corrected weight adjustments for $\hat{Y}(1/2)_{bc}$ are $\tilde{m}_i^{A} = 19/18$ for i classified in a and $\tilde{m}_i^{A} = 1/2$ for i classified in ab; for pattern (c), the bias-corrected weight adjustments are 17/16 and 7/16, respectively. The single frame estimator is omitted from these tables since it is the same as either $\hat{Y}(1/2)$ or $\hat{Y}(2/3)$; the single frame estimator raked to the population totals N_A and N_B is denoted by $\hat{Y}_{\text{SF, rake}}$. Tables 1 and 2 display results for $n_A = 100, n_B = 100,$ $N_a = N_{ab} = N_b = 25{,}000,$ and a simple random sample from frame A; Tables 3 and 4 give results for $n_A = 200,$ $n_B = 100,$ $N_a = N_{ab} = N_b = 25{,}000,$ and a cluster sample from frame A. The general patterns of results are similar for the other simulations and are not shown here.

First, consider the fixed weight estimators. The bias-corrected estimators reduce the bias as expected; in all cases studied with misclassification, the empirical bias from the bias-corrected estimators was less than 200 in absolute value, which is within the margin of error. Although the standard deviation for the bias-corrected estimators is higher than for the uncorrected estimators, in most cases the mean squared errors are comparable.

The screening estimator $\hat{Y}(1)$, which discards units from frame B in domain ab, exhibits no misclassification bias when frame B is correctly classified. It also exhibits no bias in Tables 1 and 3 with frame-B misclassification pattern (d) because the observations misclassified from domain ab to domain b have mean 0; for different sets of domain means, pattern (d) does create bias. For the other cases, the screening estimator has the highest bias. For every misclassification pattern, the screening estimator has high mean squared error because data are thrown away. If the domain means are similar, then the misclassification might not result in appreciable bias but discarding observations from domain ab in $\mathcal{S}(B)$ would greatly increase the mean squared error.

Poststratifying to the domain totals when there is misclassification often increases the bias instead of decreasing it. Consider line 4 of Table 1, where 20% of the $\mathcal{S}(B)$ observations in ab are mistakenly classified into domain b. The weights of the observations that are really in domain b, with mean 2, are reduced from 500 to approximately 417, which causes the poststratified versions of $\hat{Y}(1/2)$ to be biased. The effect of poststratification on the mean squared error is mixed, and depends on whether the variance reduction achieved by poststratifying exceeds the additional bias that can be introduced. Raking to the frame totals N_A and N_B, in $\hat{Y}_{\text{SF, rake}}$, has similar effect on misclassification bias as poststratification.

For the simple random samples in Tables 1 and 2, the PML and PEL estimators often exhibit much more bias than the uncorrected fixed weight estimators. The relative contributions from the two frames for these methods depend on the estimated variances of $\hat{N}_{ab}^{A}$ and $\hat{N}_{ab}^{B}$, the domain weights depend on $\hat{N}_{ab}^{\text{PML}}$, and these two factors interact in complex ways depending on the misclassification structure. For misclassification pattern (d) in either frame, $\hat{N}_{ab}^{\text{PML}}$ is too small because observations in domain ab are misclassified; consequently, the weights for the observations in the nonoverlapping domains are too large. A poststratified version of the PML estimator shared the bias problems of the fixed weight poststratified estimators. The PEL estimator, by forcing the estimators of Y_{ab} to be equal, can worsen the bias. For example, in the simulation in line 3 of Table 1, with correct classification for frame A and pattern (c) for frame B, the PEL bias is 50% larger than the PML bias. In this case, the PEL estimator pulls the unbiased estimator $\hat{Y}_{ab}^{A}$ from $\mathcal{S}(A)$ toward the biased estimator from frame B. The optimal estimators also exhibit high bias.

Table 1
Estimated bias for dual frame misclassification, with $n_A = n_B = 100$ and a simple random sample taken from each frame. MPA and MPB refer to the misclassification patterns for frames A and B

MPA	MPB	$\hat{Y}(1/2)$	$\hat{Y}(1/2)_{\text{post1}}$	$\hat{Y}(1/2)_{\text{post2}}$	$\hat{Y}(1/2)_{bc}$	$\hat{Y}(2/3)$	$\hat{Y}(2/3)_{bc}$	$\hat{Y}(1)$	$\hat{Y}_H$	$\hat{Y}_{\text{FB}}$	$\hat{Y}_{\text{PML}}$	$\hat{Y}_{\text{PEL}}$	$\hat{Y}_{\text{SF, rake}}$
a	a	-194	-87	-87	-194	-215	-215	-258	-68	10	-121	-119	-163
a	b	-5,015	4,145	4,529	5	-6,678	17	-10,002	-5,417	1,248	2,486	1,542	2,361
a	c	-5,142	-1,118	-898	-128	-6,823	-138	-10,185	-5,413	-2,583	-1,650	-2,482	-1,690
a	d	-57	-8,430	-8,431	-47	-69	-55	-92	30	-6,576	-6,723	-6,725	-6,795
b	a	1,163	-1,238	-1,290	-82	748	-82	-82	1,355	-2,376	-2,631	-2,551	-2,704
b	b	-3,724	3,040	3,264	43	-5,784	65	-9,905	-3,967	-920	-30	-850	-100
b	c	-3,882	-2,192	-2,187	-124	-5,977	-136	-10,167	-3,954	-4,319	-3,821	-4,477	-3,853
b	d	1,322	-9,445	-9,621	92	917	104	108	1,600	-8,219	-8,720	-8,531	-8,879
c	a	1,366	1,315	1,312	123	969	140	174	1,530	1,529	1,325	1,355	1,276
c	b	-3,729	5,456	5,948	51	-5,801	64	-9,945	-4,216	2,096	3,500	2,391	3,355
c	c	-3,797	235	512	-15	-5,868	2	-10,011	-4,089	-1,377	-417	-1,318	-466
c	d	1,285	-7,072	-7,212	56	873	60	48	1,535	-4,665	-5,131	-4,976	-5,222
d	a	-120	2,134	2,134	-111	-132	-126	-155	32	3,710	3,535	3,538	3,470
d	b	-4,979	6,497	7,086	34	-6,620	65	-9,901	-5,599	4,339	5,928	4,788	5,697
d	c	-5,152	1,174	1,644	-137	-6,835	-152	-10,200	-5,622	310	1,626	578	1,540
d	d	90	-5,999	-5,998	107	98	119	114	193	-2,964	-3,116	-3,120	-3,155

Table 2
Estimated $\sqrt{\text{MSE}}$ for dual frame misclassification, with $n_A = n_B = 100$ and a simple random sample taken from each frame. MPA and MPB refer to the misclassification patterns for frames A and B

MPA	MPB	$\hat{Y}(1/2)$	$\hat{Y}(1/2)_{\text{post1}}$	$\hat{Y}(1/2)_{\text{post2}}$	$\hat{Y}(1/2)_{bc}$	$\hat{Y}(2/3)$	$\hat{Y}(2/3)_{bc}$	$\hat{Y}(1)$	$\hat{Y}_H$	$\hat{Y}_{\text{FB}}$	$\hat{Y}_{\text{PML}}$	$\hat{Y}_{\text{PEL}}$	$\hat{Y}_{\text{SF, rake}}$
a	a	9,646	7,917	7,910	9,646	9,729	9,729	10,304	9,677	8,151	8,081	8,115	8,075
a	b	10,602	9,351	9,531	9,926	11,531	10,197	14,181	11,157	8,212	8,377	8,198	8,311
a	c	10,779	8,622	8,603	10,071	11,715	10,402	14,376	11,243	8,817	8,514	8,720	8,508
a	d	9,789	11,719	11,704	9,674	9,884	9,795	10,432	9,819	10,979	10,978	11,003	11,007
b	a	9,623	8,182	8,185	9,718	9,686	9,766	10,307	9,780	8,446	8,447	8,444	8,459
b	b	9,955	9,054	9,137	9,995	10,949	10,212	14,069	10,489	8,074	7,913	7,995	7,898
b	c	10,146	9,014	9,014	10,160	11,197	10,489	14,404	10,616	9,443	9,108	9,448	9,114
b	d	9,868	12,600	12,716	9,826	9,952	9,927	10,567	10,023	12,063	12,284	12,188	12,371
c	a	9,843	8,185	8,180	9,887	9,853	9,877	10,341	9,991	8,516	8,417	8,442	8,402
c	b	10,049	10,113	10,396	10,039	11,029	10,229	14,127	10,662	8,520	8,863	8,529	8,778
c	c	10,247	8,701	8,718	10,254	11,233	10,534	14,306	10,799	8,762	8,527	8,669	8,516
c	d	10,021	10,861	10,936	9,966	10,068	10,016	10,579	10,177	10,113	10,211	10,168	10,240
d	a	9,795	8,127	8,121	9,734	9,845	9,788	10,343	9,829	9,158	9,024	9,042	8,991
d	b	10,718	10,601	10,970	10,001	11,602	10,258	14,149	11,358	9,461	10,157	9,595	9,986
d	c	10,847	8,558	8,650	10,099	11,769	10,426	14,387	11,424	8,674	8,707	8,608	8,664
d	d	9,945	10,070	10,057	9,778	10,019	9,885	10,510	9,986	9,458	9,412	9,449	9,417

When a cluster sample is taken from frame A, as in Tables 3 and 4, the bias patterns are similar. When there is no misclassification, the MSEs of the optimal and PML estimators are smaller than that of $\hat{Y}(2/3)$ because they account for the survey design. With misclassification, though, the MSE advantage is reduced because of the increased bias.

To study misclassification with a three-frame survey, we selected simple random samples from each frame, and had correct classifications for frames B and C. Table 5 shows results for a simulation with three frames and a simple random sample of size 200 from each frame. The population was generated with $N_d = 10{,}000$ in each domain and domain means $\mu_a = 1$, $\mu_{ab} = 2$, $\mu_{ac} = 3$, $\mu_{abc} = 4$, $\mu_b = 5$, $\mu_{bc} = 6$, $\mu_c = 7$. In this simulation, frames B and C are correctly classified, and the misclassification patterns for frame A are given in the table. We also studied other domain means, population domain sizes, and sample sizes using a factorial design; results for the other settings showed a similar pattern and are not shown here. The multiplicity estimator $\hat{Y}_{\text{ave}}$, with $m_i = 1$ for $i \in \{a, b, c\}$, $m_i = 1/2$ for $i \in \{ab, ac, bc\}$, and $m_i = 1/3$ for $i \in abc$, is optimal when there is no misclassification, and it equals the unraked single frame estimator. The other fixed weight estimators studied are $\hat{Y}_{2/3}$, with $m^{(A,a)} = m^{(B,b)} = m^{(C,c)} = 1$, $m^{(A,ab)} = m^{(A,ac)} = m^{(A,abc)} = 2/3$, $m^{(B,ab)} = m^{(C,ac)} = 1/3$, and $m^{(B,abc)} = m^{(C,abc)} = 1/6$, and the screening estimator $\hat{Y}_{\text{scr}}$, with $m^{(A,a)} = m^{(B,b)} = m^{(C,c)} = m^{(A,ab)} = m^{(A,ac)} = m^{(A,abc)} = m^{(B,bc)} = 1$.

As with the two-frame study, the bias-corrected estimators are approximately unbiased. The screening estimator is also approximately unbiased since only $\mathcal{S}(A)$ is misclassified. The other estimators all exhibit substantial bias with at least some of the misclassification patterns. For the simulation settings in Table 5, the poststratified, single frame raking, Hartley, and PML estimators exhibit large bias but nevertheless have smaller mean squared error than the fixed weight and bias-corrected estimators; this MSE ordering does not hold in some of the other simulation settings.

Mecatti (2007) and Rao and Wu (2010) argued that the fixed weight multiplicity estimator $\hat{Y}_{\text{ave}}$ is unbiased if the only misclassification is among domains that belong to the same number of frames. Misclassifying observations from domain ab to domain ac (pattern c) results in no bias because the weight adjustment in both domains is $1/2$. In practice, though, one would expect pattern (c), with two errors in domain membership (not reporting membership in frame B and erroneously reporting membership in frame C), to be less likely to occur in practice than misclassifying an observation in ab as either a or abc; $\hat{Y}_{\text{ave}}$ can be very sensitive to the latter forms of misclassification. Although a fixed weight estimator is insensitive to misclassification among domains in which the weight adjustments are equal, in these simulations every fixed weight estimator exhibits significant bias for at least some misclassification patterns.

Tables 1 to 5 show that each estimator from Section 2 can exhibit severe bias from domain misclassification. We recommend that the possible extent of domain misclassification be studied during the survey pretesting phase, so that this information can be used in the survey design. If misclassification probabilities are known accurately, then it may be possible to choose a fixed weight estimator that is insensitive to the presumed form of misclassification. When a misclassification-robust estimator cannot be found or when it is inefficient, the fixed weight estimators can be adjusted to reduce the bias. It should be noted that the bias-corrected weights proposed in Section 6.1 are sensitive to the input misclassification probabilities. They also do not account for other nonsampling errors such as nonresponse; applying the misclassification weight adjustments in Section 6.1 followed by the nonresponse weight adjustments described in Brick *et al.* (2011) may result in final weights that correct neither for misclassification nor for nonresponse. If domain misclassification and nonresponse are both present, weight adjustments are needed that deal with both problems simultaneously.

Table 3
Estimated bias for dual frame misclassification, with n_A = 200, n_B = 100, a cluster sample taken from frame A and a simple random sample taken from frame B. MPA and MPB refer to the misclassification patterns for frames A and B

MPA	MPB	$\hat{Y}(1/2)$	$\hat{Y}(1/2)_{\text{post1}}$	$\hat{Y}(1/2)_{\text{post2}}$	$\hat{Y}(1/2)_{bc}$	$\hat{Y}(2/3)$	$\hat{Y}(2/3)_{bc}$	$\hat{Y}(1)$	$\hat{Y}_H$	$\hat{Y}_{\text{FB}}$	$\hat{Y}_{\text{PML}}$	$\hat{Y}_{\text{PEL}}$	$\hat{Y}_{\text{SF, rake}}$
a	a	-148	-142	-139	-148	-155	-155	-170	-312	63	-119	-172	-184
a	b	-5,090	4,199	4,599	-72	-6,774	-84	-10,144	-4,976	1,210	3,615	2,181	1,025
a	c	-5,069	-1,088	-851	-72	-6,759	-96	-10,139	-4,800	-1,994	177	-1,136	-3,216
a	d	-39	-8,379	-8,383	-35	-63	-58	-111	-237	-5,757	-5,909	-5,961	-6,996
b	a	1,168	-1,221	-1,258	-79	768	-63	-32	1,395	-1,690	-1,663	-2,514	-3,170
b	b	-3,716	2,979	3,236	60	-5,784	79	-9,918	-2,815	-86	1,776	-346	-2,087
b	c	-3,704	-2,108	-2,074	73	-5,771	92	-9,905	-2,561	-2,970	-1,410	-3,267	-5,814
b	d	1,317	-9455	-9,610	95	926	123	144	1,609	-7,285	-7,317	-7,938	-9,498
c	a	1,179	1,281	1,304	-66	772	-58	-41	1,486	1,831	1,652	943	840
c	b	-3,879	5,545	6,087	-118	-5,971	-126	-10,156	-2,972	3,532	4,597	2,405	1,683
c	c	-3,811	318	636	-44	-5,893	-42	-10,058	-2,671	110	1,128	-784	-2,328
c	d	1,423	-6,858	-6,973	191	1,022	206	220	1,824	-4,328	-4,014	-4,516	-5,624
d	a	-33	2,282	2,290	-28	-35	-32	-40	-148	3,627	3,138	3,103	3,728
d	b	-4,974	6,514	7,123	46	-6,660	30	-10,033	-4,863	4,768	6,274	4,742	4,549
d	c	-4,951	1,412	1,883	80	-6,621	84	-9,961	-4,682	1,357	2,863	1,451	388
d	d	42	-5,987	-5,991	53	40	52	37	-126	-2,899	-2,780	-2,791	-3,317

Table 4
Estimated $\sqrt{\text{MSE}}$ for dual frame misclassification, with n_A = 200, n_B = 100, a cluster sample taken from frame A and a simple random sample taken from frame B. MPA and MPB refer to the misclassification patterns for frames A and B

MPA	MPB	$\hat{Y}(1/2)$	$\hat{Y}(1/2)_{\text{post1}}$	$\hat{Y}(1/2)_{\text{post2}}$	$\hat{Y}(1/2)_{bc}$	$\hat{Y}(2/3)$	$\hat{Y}(2/3)_{bc}$	$\hat{Y}(1)$	$\hat{Y}_H$	$\hat{Y}_{\text{FB}}$	$\hat{Y}_{\text{PML}}$	$\hat{Y}_{\text{PEL}}$	$\hat{Y}_{\text{SF, rake}}$
a	a	10,916	8,912	8,899	10,916	11,092	11,092	11,879	10,975	9,250	9,155	10,109	9,418
a	b	11,786	10,186	10,324	11,157	12,743	11,503	15,463	12,253	8,906	9,391	10,123	9,231
a	c	11,983	9,575	9,537	11,409	12,922	11,814	15,600	12,395	9,575	9,279	10,391	10,039
a	d	11,042	12,357	12,375	10,941	11,250	11,173	12,056	11,051	11,591	11,605	12,229	12,053
b	a	10,698	9,133	9,154	10,872	10,921	11,049	11,875	10,823	9,255	9,151	10,195	9,766
b	b	10,957	9,803	9,867	11,071	12,033	11,413	15,262	11,215	8,681	8,748	9,610	9,182
b	c	11,115	9,860	9,846	11,272	12,172	11,675	15,361	11,306	9,721	9,252	10,558	10,970
b	d	10,988	13,269	13,408	11,046	11,222	11,262	12,143	11,084	12,484	12,347	13,279	13,598
c	a	10,995	9,090	9,073	11,106	11,187	11,254	12,028	11,125	9,309	9,190	9,798	9,389
c	b	11,104	10,779	11,015	11,090	12,162	11,380	15,348	11,430	9,450	9,724	9,754	9,144
c	c	11,155	9,425	9,400	11,189	12,234	11,600	15,424	11,389	9,219	9,064	9,868	9,658
c	d	10,922	11,328	11,421	10,896	11,121	11,091	11,929	11,017	10,759	10,456	11,151	11,182
d	a	11,011	9,080	9,045	10,920	11,181	11,103	11,913	11,041	9,873	9,579	10,375	10,135
d	b	11,838	11,357	11,669	11,164	12,723	11,453	15,299	12,337	10,258	10,848	11,009	10,403
d	c	11,804	9,334	9,371	11,159	12,707	11,548	15,298	12,224	9,349	9,507	10,102	9,442
d	d	11,179	10,839	10,854	10,989	11,355	11,199	12,059	11,195	10,440	10,302	10,916	10,519

Table 5
Estimated bias and $\sqrt{\text{MSE}}$ for misclassification in a 3-frame survey, with $n_A = n_B = n_C = 200$ and a simple random sample taken from each frame. MPA refers to the misclassification patterns for frame A. Pattern (a) has no misclassification; (b) $\phi^A_{aa} = 0.8$, $\phi^A_{a,ab} = 0.1$, $\phi_{a,abc} = 0.1$, $\phi^A_{ab,ab} = 1$, $\phi^A_{ac,ac} = 1$, $\phi^A_{abc,abc} = 1$; (c) $\phi^A_{aa} = 1$, $\phi^A_{ab,ab} = 0.9$, $\phi^A_{ab,ac} = 0.1$, $\phi^A_{ac,ac} = 1$, $\phi^A_{abc,abc} = 1$; (d) $\phi^A_{aa} = 1$, $\phi^A_{ab,ab} = 0.9$, $\phi^A_{ab,abc} = 0.1$, $\phi^A_{ac,ac} = 1$, $\phi^A_{abc,abc} = 1$; (e) $\phi^A_{aa} = 1$, $\phi^A_{ab,ab} = 0.8$, $\phi^A_{ab,a} = 0.1$, $\phi_{ab,abc} = 0.1$, $\phi^A_{ac,ac} = 1$, $\phi^A_{abc,abc} = 1$

	MPA	$\hat{Y}_{\text{ave}}$	$\hat{Y}_{\text{ave, post1}}$	$\hat{Y}_{\text{ave, post2}}$	$\hat{Y}_{\text{ave}, bc}$	$\hat{Y}_{2/3}$	$\hat{Y}_{2/3,\, bc}$	$\hat{Y}_{\text{scr}}$	$\hat{Y}_H$	$\hat{Y}_{\text{PML}}$	$\hat{Y}_{\text{SF, rake}}$
	a	-8	31	28	-8	5	5	20	9	-26	-208
	b	-938	-1,409	-1,478	57	-586	77	107	-2,039	-5,676	-5,624
bias	c	-26	-485	-508	-26	6	6	6	-324	-825	-957
	d	-231	-514	-557	104	108	108	85	-326	-1,321	-1,438
	e	704	287	247	34	697	27	-4	1,488	1,420	1,193
	a	9,003	4,419	4,410	9,003	10,013	10,013	13,108	7,990	7,281	7,293
	b	8,961	4,711	4,730	8,955	9,952	9,953	13,092	8,085	9,107	9,074
$\sqrt{\text{MSE}}$	c	9,119	4,432	4,422	9,119	10,140	10,140	13,238	8,112	7,396	7,422
	d	8,894	4,405	4,405	8,893	9,874	9,874	12,919	7,957	7,414	7,433
	e	9,088	4,438	4,424	9,059	10,071	10,046	13,180	8,254	7,621	7,581

7. Design issues

As discussed in Section 1, multiple frame designs can give better coverage and precision than a single frame survey with equivalent cost. The design problem is more complex than with a single frame survey, though, since a design that is optimal for frame A and frame B separately may not be optimal for the combined sample. Similarly, a design that is optimal when estimator $\hat{Y}(1/2)$ is used may not be optimal for $\hat{Y}_{\text{PML}}$.

Hartley (1962, 1974) derived optimal designs for the estimator $\hat{Y}(\hat{\theta}_H)$ when a simple random sample is taken in each frame. The optimal sample sizes n_A and n_B depend on the relative costs of sampling from the two frames, and on the means and variances of the response variable within the domains. Cochran (1977, pages 144-145) described the dual frame survey in Figure 1 in his chapter on stratified sampling. In this situation, N_a and N_{ab} may be known, especially if frame B is a list frame. Domains a and ab are treated as strata; there is one sample from stratum a and two independent samples from stratum ab. The design problem may be approached as a stratified sample design.

In general, the optimal design is a function of sampling variances and nonsampling errors in each frame, as well as of the estimator chosen. Biemer (1984) and Lepkowski and Groves (1986) discussed designs for the situation in Figure 1 when a stratified multistage sample is taken from each frame, using the Hartley estimator $\hat{Y}(\hat{\theta}_H)$. Lepkowski and Groves (1986) considered interviewer variability and mode bias as well as sampling error when assessing the precision of various designs; frames with less mode bias are allotted higher sample sizes. Brick (2010) derived optimal allocations in the presence of nonresponse, and found that considering the nonresponse when allocating resources to the two frames can greatly increase efficiency in both screening and overlap dual frame surveys.

One of the advantages of a multiple frame design is its flexibility; it is well suited for a modular approach to survey design. In some situations, it may be practical to take an initial sample from the general population (frame A in Figure 4). The design of the samples from frames B and C, corresponding to subpopulations of interest, can then be determined using information in the frame-A sample. For example, if the frame-A sample yields too few engineers, the sample size from an engineering society membership list frame can be correspondingly increased.

Rao (2003) suggested using multiple frame surveys to improve the accuracy of small area estimates in subgroups of interest. In this application, supplemental surveys can be taken in frames with high concentrations of subgroups of interest. As research needs change, resources can be re-allocated among the supplemental surveys without changing the main survey design. A crime victimization survey that uses a national area frame may be supplemented by local victimization surveys; as victimization patterns change, the local surveys can have different sample sizes or be moved to other geographic regions.

Most survey designs attempt to achieve efficiency for the important responses, but in some situations a design that is efficient for one response is inefficient for others. For a survey in which each of four responses of interest was highly correlated with one of the possible stratification variables (but not necessarily correlated with the other stratification variables), Skinner, Holmes and Holt (1994) used a multiple frame survey with four independent stratified samples drawn from a common sampling frame. Each sample was stratified using the stratification variable that was correlated with one of the responses of interest, and so was highly efficient for that response. In estimation, information from all four samples was combined.

Multiple frame surveys can also be used in conjunction with sequential or adaptive sampling methods to improve yield of a rare or hard-to-reach population such as recent immigrants. For example, a stratified multistage sampling design might be employed for frame A, while an adaptive cluster sampling design (Thompson 2002) might be used for frame B. Domain estimates can be calculated separately for the two designs, and then combined using methods in Section 2. In this situation, frames A and B may completely overlap, so that domain misclassification will not be an issue.

8. Conclusions

In this paper, we have summarized some of the issues involved in using multiple frame methods for U.S. household surveys. Multiple frame designs have great potential for improving efficiency of data collection in household surveys. They can improve coverage by combining incomplete frames, improve the accuracy of estimates for subgroups or rare populations, and increase the flexibility and responsiveness of federal data collection. Multiple frame surveys can facilitate sampling hard-to-reach populations such as recent immigrants or households with infants; a general population survey can be combined with an adaptive sample design or a list frame of births.

In many cases, multiple frame surveys can provide more accurate estimates of population quantities without increasing data collection costs, but the design and estimator must be chosen carefully to realize these savings. A multiple frame survey, like other surveys, may have nonresponse, mode effects, and measurement errors. In addition, unless all of the frames consist of the entire population, multiple frame survey estimators can be sensitive to domain misclassification. One correction for misclassification was given in Section 6, but more research is needed on these challenges. Effects of domain misclassification, nonresponse, and mode bias may be confounded. A designed experiment may help disentangle these effects. We are currently studying the relation among these three types of nonsampling errors. Each form of nonsampling error affects the accuracy of multiple frame estimators, and anticipated nonsampling errors need to be incorporated in an optimal design.

Acknowledgements

This research was partially supported by the National Science Foundation under grants SES-0604373 and DRL-0909630. Some of the material in this paper was presented at the 2010 annual meeting of the Italian Statistical Society and published in the proceedings for that conference. The author is grateful to the reviewers for their helpful comments.

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.

Biemer, P.P. (1984). Methodology for optimal dual frame sample design. Bureau of the Census SRD Research Report CENSUS/SRD/RR-84/07.

Brick, J.M. (2010). Dual frame landline and cell phone surveys. Paper presented at the annual meeting of the Statistical Society of Canada, Québec City.

Brick, J.M., Cervantes, I.F., Lee, S. and Norman, G. (2011). Nonsampling errors in dual frame telephone surveys. *Survey Methodology*, 37, 1, 1-12.

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

Chambers, R., Chipperfield, J., Davis, W. and Kovačević, M. (2008). Inference based on estimating equations and probability-linked data. University of Wollongong Centre for Statistical & Survey Methodology Working Paper 18-09.

Citro, C.F., and Kalton, G., Eds. (2007). *Using the American Community Survey: Benefits and Challenges.* Washington, D.C.: National Academies Press.

Clark, J., Winglee, M. and Liu, B. (2007). Handling imperfect overlap determination in a dual-frame survey. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 3233-3238.

Cochran, W.G. (1977). *Sampling Techniques*, 3rd Ed. New York: John Wiley & Sons, Inc.

de Leeuw, E. (2008). Choosing the method of data collection. In *International Handbook of Survey Methodology*, (Eds., E. de Leeuw, J. Hox and D. Dillman). New York: Lawrence Erlbaum, 113-135.

de Leeuw, E., Hox, J. and Dillman, D. (2008). Mixed-mode surveys: When and why. In *International Handbook of Survey Methodology*, (Eds., E. de Leeuw, J. Hox and D. Dillman). New York: Lawrence Erlbaum, 299-316.

Demnati, A., Rao, J.N.K., Hidiroglou, M.A. and Tambay, J.-L. (2007). On the allocation and estimation for dual frame survey data. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 2938-2945.

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

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

González-Villalobos, A., and Wallace, M.A. (1996). *Multiple Frame Agriculture Surveys*, Rome: Food and Agriculture Organization of the United Nations. Vols. 1 and 2.

Haines, D.E., and Pollock, K.H. (1998). Combining multiple frames to estimate population size and totals. *Survey Methodology*, 24, 79-88.

Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953). *Sample Survey Methods and Theory*. New York: John Wiley & Sons, Inc. Volume 1.

Hartley, H.O. (1962). Multiple frame surveys. *Proceedings of the Social Statistics Section*, American Statistical Association, 203-206.

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

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.

Iachan, R., and Dennis, M.L. (1993). A multiple frame approach to sampling the homeless and transient population. *Journal of Official Statistics*, 9, 747-764.

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

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

Kott, P.S., Amrhein, J.F. and Hicks, S.D. (1998). Sampling and estimation from multiple list frames. *Survey Methodology*, 24, 3-9.

Lepkowski, J.M., and Groves, R.M. (1986). A mean squared error model for multiple frame, mixed mode survey design. *Journal of the American Statistical Association*, 81, 930-937.

Lesser, V.M., and Kalsbeek, W.D. (1999). Nonsampling errors in environmental surveys. *Journal of Agricultural, Biological, and Environmental Statistics*, 4, 473-488.

Lohr, S.L. (2007). Recent developments in multiple frame surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 3257-3264.

Lohr, S.L. (2009). Multiple frame surveys. In *Handbook of Statistics, Sample Surveys: Design, Methods and Applications*, (Eds., D. Pfeffermann and C.R. Rao). Amsterdam: North Holland, Vol. 29A, 71-88.

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

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

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

Lu, Y., and Lohr, S.L. (2010). Gross flow estimation in dual frame surveys. *Survey Methodology*, 36, 13-22.

Mecatti, F. (2007). A single frame multiplicity estimator for multiple frame surveys. *Survey Methodology*, 33, 151-157.

National Science Foundation (2003). *SESTAT: Design and Methodology*, http://srsstats.sbe.nsf.gov/docs/techinfo.html.

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

Rao, J.N.K., and Wu, C. (2010). Pseudo-empirical likelihood inference for dual frame surveys. *Journal of the American Statistical Association*, 105, 1494-1503.

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

Skinner, C.J., Holmes, D.J. and Holt, D. (1994). Multiple frame sampling for multivariate stratification. *International Statistical Review*, 62, 333-347.

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.

Thompson, S.K. (2002). *Sampling Techniques*, 2nd Ed. New York: John Wiley & Sons, Inc.

Vannieuwenhuyze, J., Loosveldt, G. and Molenberghs, G. (2011). A method for evaluating mode effects in mixed-mode surveys. *Public Opinion Quarterly*, 74, 1027-1045.

Survey Methodology, December 2011
Vol. 37, No. 2, pp. 215-226
Statistics Canada, Catalogue No. 12-001-X

Ten years of balanced sampling with the cube method: An appraisal

Yves Tillé [1]

Abstract

This paper presents a review and assessment of the use of balanced sampling by means of the cube method. After defining the notion of balanced sample and balanced sampling, a short history of the concept of balancing is presented. The theory of the cube method is briefly presented. Emphasis is placed on the practical problems posed by balanced sampling: the interest of the method with respect to other sampling methods and calibration, the field of application, the accuracy of balancing, the choice of auxiliary variables and ways to implement the method.

Key Words: Sampling; Balancing; Horvitz-Thompson estimator.

1. Introduction

While the idea of balanced sampling has been around since the early days of survey statistic development, applying the concept has been difficult because almost all the proposed methods have either been enumerative or rejective and required considerable computation time. The algorithm of the cube method was proposed in 1998 by Deville and Tillé, and a first implementation was written by three students of the École Nationale de la Statistique et de l'Analyse de l'Information of Rennes in France (see Bousabaa, Lieber and Sirolli 1999). Finally, the method was published in Tillé (2001) and Deville and Tillé (2004). Since this time, several implementations have been proposed and several survey managers have used the cube method to select samples, the most important applications being the New French Census and the French Master Sample.

Our aim is to assess 10 years of development and use of balanced sampling in order to better ascertain when and how the cube method can be used to select samples of householders or establishments. After discussing the concept of balanced sample and balanced sampling in Section 2, we give a list of particular cases in Section 3. In Section 4, we briefly trace the history of this concept for both the model-based and design-based frameworks. Next, in Section 5, we provide a brief overview of the cube method, which is a class of algorithms that allows us to select randomly balanced samples with given inclusion probabilities (see Deville and Tillé 2004; Tillé 2001, 2006b). We try to present the main principles of this algorithm without giving a detailed description of the technicalities of the method. Section 6 is devoted to the principles of variance estimation in balanced sampling. Finally, in Sections 7, we discuss the interest of balanced sampling in practice and compare balanced sampling with other sampling methods and calibration. We also give a list of recent applications. This Section also deals with the accuracy of balancing, the choice of auxiliary variables and ways to implement balanced sampling. The paper ends with an exhaustive bibliographical list of references on balanced sampling and their applications.

2. Balanced sampling

2.1 Definition of a balanced sample

Consider a sample s of size n that is a subset of a finite population U of size N. A sample is said to be balanced if, for a vector of auxiliary variable $\mathbf{x}_k = (x_{k1}, \ldots, x_{kj}, \ldots, x_{kp})'$,

$$\frac{1}{n}\sum_{k\in S}\mathbf{x}_k = \frac{1}{N}\sum_{k\in U}\mathbf{x}_k, \qquad (1)$$

which means that the sample means of the x-variables match their population means.

Brewer (1999) drew a distinction between a balanced selection of samples and a random selection of samples. However, a balanced sample may be selected randomly. If a random sample S is selected randomly, then each unit of the population has an inclusion probability π_k of being selected. In this case, a random sample must satisfy the following balancing equations:

$$\sum_{k\in S}\frac{\mathbf{x}_k}{\pi_k} = \sum_{k\in U}\mathbf{x}_k. \qquad (2)$$

In other words, in a balanced sample, the total of the x-variables are estimated without error. Several authors like Cumberland and Royall (1981) and Kott (1986) would call a sample that satisfies Equation (2) a 'π-balanced sample', as opposed to a 'mean-balanced sample' defined by Equation (1). Nevertheless, in this paper, we will consider that (1) is only a particular case of (2) that occurs when $\pi_k = n/N$ or when the sample is not selected randomly. We refer to both cases as a balanced sample.

[1] Yves Tillé, University of Neuchâtel, Pierre à Mazel 7, 2000 Neuchâtel Switzerland. E-mail : yves.tille@unine.ch.

2.2 Balanced sampling design

Let $p(s)$ denote the sampling design, *i.e.*, the probability that sample s is selected, such that $p(s) = \Pr(S = s)$, where S is the random sample and $n(S)$ the size of the sample S. According to the definition of Deville and Tillé (2004), a sampling design $p(\cdot)$ is said to be *balanced* on auxiliary variables $x_1, \ldots, x_p$ if the Horvitz-Thompson estimator satisfies Equation (2). In a balanced sampling design, the inclusion probabilities are decided prior to sampling. A balanced sampling can be viewed as a kind of calibration that is directly integrated into the sampling design. The main problem is that the balancing equations (2) can rarely be exactly satisfied. We refer to this difficulty as the 'rounding problem'.

Example 1. If $N = 4, n = 2, \pi_k = 1/2,$ for all $k \in U$ and $x_1 = 0, x_2 = 1, x_3 = 2, x_4 = 4,$ then the balancing equations given in (2) becomes

$$\frac{1}{n}\sum_{k\in s} x_k = \frac{1}{N}\sum_{k\in U} x_k,$$

which is equivalent to

$$\sum_{k\in s} x_k = \frac{n}{N}\sum_{k\in U} x_k. \qquad (3)$$

Since

$$\frac{n}{N}\sum_{k\in U} x_k = \frac{2}{4}(0 + 1 + 2 + 4) = 3.5,$$

and the left side of (3) is always an integer, then an exactly balanced sample does not exist.

Indeed, sample selection is an integer problem. The cube method therefore aims to select a sample that exactly satisfies the inclusion probabilities π_k while remaining as balanced as possible.

3. Special cases of balanced sampling

3.1 Unequal probability sampling and stratification

Some well-known sampling designs are particular cases of balanced sampling:

1. Sampling with a fixed sample size is a particular case of balanced sampling. In this case, the only balancing variable is π_k. The balancing equations given in (2) become

$$\sum_{k\in S}\frac{\pi_k}{\pi_k} = \sum_{k\in S} 1 = \sum_{k\in U}\pi_k,$$

which means that the sample size must be fixed.

2. Stratification is a particular case of balanced sampling. Suppose that the population is partitioned in H strata $U_h, h = 1, \ldots, H,$ of sizes $N_h, h = 1, \ldots, H,$ and that a sample is selected in each stratum by means of simple random sampling without replacement with fixed sample size $n_h, h = 1, \ldots, H.$ In this case, the balancing variables are the indicator variables of the strata

$$\delta_{kh} = \begin{cases} 1 & \text{if } k \in U_h \\ 0 & \text{otherwise.}\end{cases}$$

Under a stratified design, the Horvitz-Thompson estimators of the sizes of the strata exactly equal the sizes of the strata, which is a property of balancing on the indicator variables of the strata. Indeed, since the inclusion probabilities in stratum h are $\pi_k = n_h / N_h, k \in U_h,$ the balancing equations become

$$\sum_{k\in S}\frac{N_h \delta_{kh}}{n_h} = \sum_{k\in U}\delta_{kh} = N_h, h = 1, \ldots, H,$$

and are exactly satisfied.

These two designs are well known and are commonly applied in official statistics in order to reduce variance. The more general concept of balancing allows more freedom to choose the most appropriate balancing variables that will improve the accuracy of the estimators.

3.2 Overlapping strata

Constructing a stratified sampling design is often a difficult exercise. Statisticians often try to stratify using several qualitative variables. However, in most cases, crossing all of the strata of all the variables will cause the cells to become too small for a sample to be selected in each cell. In the context of calibration, statisticians generally calibrate on marginal totals and not on all the cells contained in a contingency table. Since a balanced sampling can be viewed as a kind of calibration that is directly integrated in the sampling design, one would also like to balance using only marginal totals. Nevertheless, the usual theory of stratification does not allow overlapping strata since the stratification must be a partition of the population. Now, the cube method enables us to directly balance on totals of overlapping strata by simply using the indicators of the strata as balancing variables.

3.3 Balancing on a constant

Another interesting special case of balanced sampling occurs when a constant is used as a balancing variable. If $\mathbf{x}_k = 1$ for all $k \in U,$ the balancing equations become

$$\sum_{k\in S}\frac{1}{\pi_k} = \sum_{k\in U} 1 = N.$$

Actually,

$$\sum_{k \in S} \frac{1}{\pi_k}$$

is the Horvitz-Thompson estimator of N. This means that, if a constant is used as a balancing variable, the estimated population size matches the known size N, which is far from being a given when the statistical units are selected with unequal inclusion probabilities.

4. History of the concept of balancing and existing methods

The idea of balanced sampling is very old and is linked to the vague concept of representativeness that was already used by Kiaer (1896, 1899, 1903, 1905). The first paper dedicated to the selection of a balanced sample is due to Gini (1928) and Gini and Galvani (1929) who selected a sample of 29 from the 214 Italian districts in order to match several population totals. Both Neyman (1952) and Yates (1960) condemned the paper of Gini and Galvani essentially because this sample was not randomly selected (see Langel and Tillé 2010). The first methods for selecting a random balanced sample were proposed by Yates (1946) and Thionet (1953), but these methods were rejective in the sense that they involved selecting samples or changing units randomly in the sample until a balanced enough sample was obtained.

In the model-based framework, Royall (1976a, b) advocated the use of balanced sampling in order to reach the optimal strategy and to protect against mis-specification of the model. (see also Royall and Pfeffermann 1982; Kott 1986; Cumberland and Royall 1988; Royall 1988; Tirari 2006; Nedyalkova and Tillé 2009). While several methods for selecting a balanced sample are presented in the book of Valliant, Dorfman and Royall (2000), these methods do not necessarily specify the inclusion probabilities of the sample. In the model-based framework, it is important to have a balanced sample. However, this sample does not always need to be randomly selected.

Hájek (1981) also advocated the use of balanced sampling. For Hájek, a balanced sampling is a particular case of representative strategy, a strategy being a couple made of a sampling design and an estimator. A representative strategy is a strategy that estimates the totals of auxiliary variables without error. In this sense, a balanced sampling design with the Horvitz-Thompson estimator is a representative strategy. Hájek (1981) proposes a rejective procedure that consists of selecting a sequence of samples until a balanced one is obtained. Rejective procedures have two drawbacks: if several balancing variables are used, the procedure can be very slow; secondly, the inclusion probabilities of rejective designs are not the same as the original design. The inclusion probabilities of statistical units that are close to the population means are increased to the detriment of the units that are far from the center (see for instance the simulations of Legg and Yu 2010).

Another method of selection consists of enumerating all the possible samples, and then constructing a sampling design only to select the samples that are adequately balanced. Such a design can be constructed by using linear programming. This technique was applied by Ardilly (1991) to select the primary units of the French master sample. Nevertheless, this method can only be applied on small population sizes because of the combinatory explosion of the number of samples when the size of the population is large.

Deville, Grosbras and Roth (1988) and Deville (1992) proposed multivariate methods for balanced sampling with equal inclusion probabilities. Hedayat and Majumdar (1995) have proposed the adaptation of an experimental design technique that would enable a balanced sampling design to be constructed. Again, this technique is restricted to equal inclusion probabilities. Finally, the cube method was proposed by Deville and Tillé (2004). This method is general in the sense that the inclusion probabilities are exactly satisfied, that these probabilities may be equal or unequal and that the sample is as balanced as possible.

Fuller (2009) studied a rejective procedure by fixing a tolerance interval outside of which the sample is rejected and proposed an estimator of variance. Even if the inclusion probabilities are changed with a rejective procedure, Fuller (2009) shows that efficient estimates are obtained by using the inclusion probabilities of the original design. Using a set of simulations, Legg and Yu (2010) compared this rejective procedure to the cube method and showed that both methods perform equally. Finally, Dudoignon and Vanheuverzwyn (2006) proposed a fast method of balanced sampling for marginal totals, whereas Périé (2008) proposed a method based on permanent random numbers that provides a balanced sample. With the Périé (2008) method, the inclusion probabilities are only approximately satisfied.

5. The cube method

5.1 Main ideas

The cube method (see Deville and Tillé 2004; Tillé 2001, 2006a, b; Ardilly 2006) is a class of sampling algorithms that selects a balanced sample and exactly satisfies a set of given inclusion probabilities. The cube method is an extension of the splitting method that was developed by Deville and Tillé (1998). It is based on a random transformation of the vector of inclusion probabilities until a sample is obtained such that:

(i) the inclusion probabilities are exactly satisfied,

(ii) the balancing equations are satisfied to the furthest extent possible.

The name of the method comes from the geometric representation of a sampling design. Indeed, a sample may be represented by a vector of samples indicators:

$$\mathbf{s} = (I[1 \in s] \ldots I[k \in s] \ldots I[N \in s])',$$

where $I[k \in s]$ takes value 1 if $k \in s$ and 0 if not. A sample may thus be viewed as a vertex of an N-cube as showed in Figure 1.

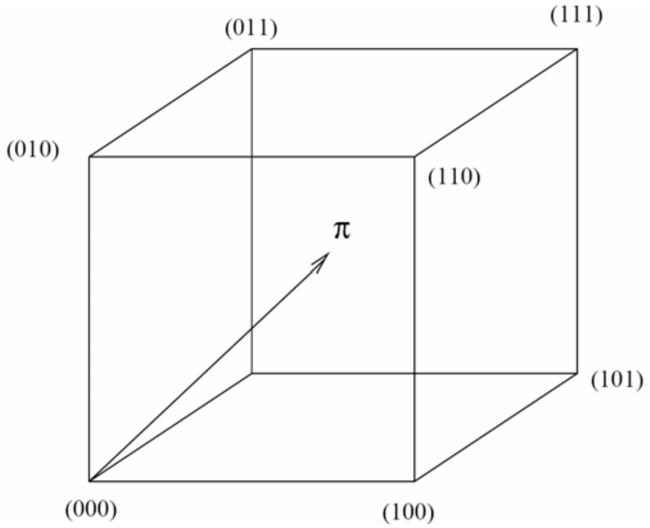


Figure 1 Possible samples in a population of size $N = 3$

Let us also define

$$\mathrm{E}(\mathbf{s}) = \sum_{s \in S} p(\mathbf{s})\mathbf{s} = \boldsymbol{\pi},$$

where $\boldsymbol{\pi} = [\pi_k]$ is the vector of inclusion probabilities. The balancing equations

$$\sum_{k \in S} \frac{\mathbf{x}_k}{\pi_k} = \sum_{k \in U} \mathbf{x}_k,$$

may also be written

$$\sum_{k \in U} \check{\mathbf{x}}_k \, s_k = \sum_{k \in U} \check{\mathbf{x}}_k \, \pi_k, \qquad (4)$$

where $s_k \in \{0, 1\}$ and $\check{\mathbf{x}}_k = \mathbf{x}_k / \pi_k, \, k \in U.$ Expression (4) is a system of equations with unknowns values s_k that define an affine subspace in $\mathbb{R}^N$ of dimension $N - p$ denoted by Q, where

$$Q = \left\{ \mathbf{u} \in \mathbb{R}^N \,|\, \sum_{k \in U} \check{\mathbf{x}}_k \, u_k = \sum_{k \in U} \mathbf{x}_k \right\}.$$

The problem of selecting a balanced sample may thus be reformulated. A balanced sampling design consists of choosing a vertex of the N-cube (a sample) that remains on the linear sub-space Q. Figures 2 and 3 respectively show two examples: the first one is a constraint of fixed sample size and the second one is a constraint that generates a rounding problem.

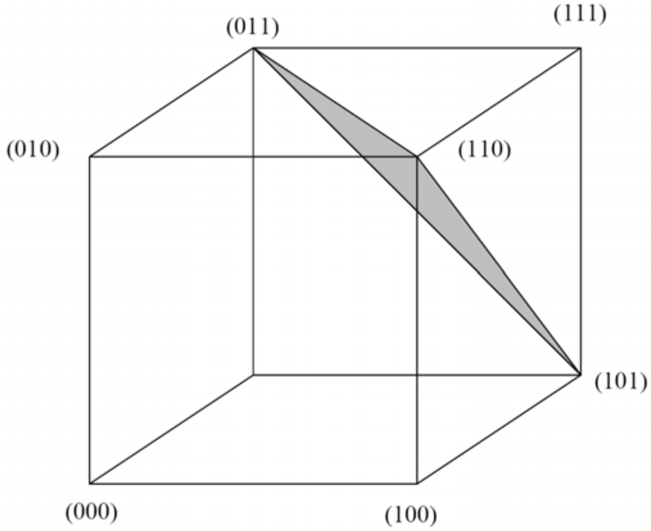


Figure 2 Possible samples in a population of size $N = 3$ with a constraint of fixed sample size $n = 2$

The Cube method (Deville and Tillé 2004) is divided into two phases: the flight phase and the landing phase. The flight phase is a random walk that begins at the vector of inclusion probabilities and remains in the intersection of the cube and the constraint subspace. This random walk stops at a vertex of the intersection of the cube and the constraint subspace. At the end of the flight phase, if a sample is not obtained, the landing phase entails in selecting a sample that is as close as possible to the constraint subspace.

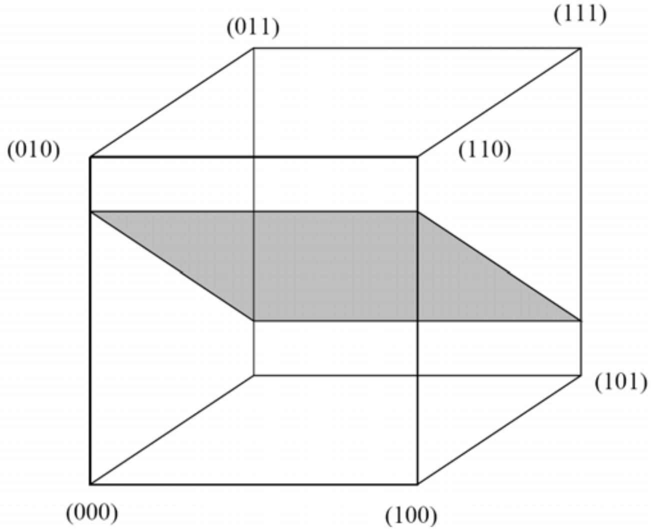


Figure 3 Possible samples in a population of size $N = 3$ with a constraint and a rounding problem

Example 2. If the constraint is the fixed sample size, the flight phase randomly transforms a vector of inclusion probabilities into a vector of 0 and 1. At each step of the algorithm, the vector of inclusion probabilities is transformed randomly, but the sum of inclusion probabilities must remain equal to n. For instance, with $\boldsymbol{\pi} = (0.5, 0.5, 0.5, 0.5)$ and $n = 2$, we are able to obtain the following sequence of vectors:

$$\pi = \begin{pmatrix} 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \end{pmatrix} \rightarrow \begin{pmatrix} 0.6666 \\ 0.6666 \\ 0.6666 \\ 0 \end{pmatrix} \rightarrow \begin{pmatrix} 1 \\ 0.5 \\ 0.5 \\ 0 \end{pmatrix} \rightarrow \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} = \mathbf{s}.$$

The algorithm ends when all the components of the vector are equal to 0 or 1.

Example 3. If the constraint is the fixed sample size, a rounding problem appears if the sum of inclusion probabilities is not an integer. If there is a rounding problem, then some components cannot be set to zero. For instance, with $\pi = (0.5, 0.5, 0.5, 0.5, 0.5)$ and

$$\sum_{k \in U} \pi_k = 2.5,$$

we may observe the following sequence of vectors:

$$\pi = \begin{pmatrix} 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \end{pmatrix} \rightarrow \begin{pmatrix} 0.625 \\ 0 \\ 0.625 \\ 0.625 \\ 0.625 \end{pmatrix} \rightarrow \begin{pmatrix} 0.5 \\ 0 \\ 0.5 \\ 1 \\ 0.5 \end{pmatrix} \rightarrow \begin{pmatrix} 1 \\ 0 \\ 0.25 \\ 1 \\ 0.25 \end{pmatrix} \rightarrow \begin{pmatrix} 1 \\ 0 \\ 0.5 \\ 1 \\ 0 \end{pmatrix} = \pi^*.$$

In this case, the flight phase cannot end with a vector of 0 or 1 of which the sum is equal to 2.5. In this case, the flight phase ends with a vector containing one non-integer component.

5.2 The flight phase

The first step of the flight phase is presented in Figure 4 for a very specific case: the population size $N = 3$. The only balancing constraint is the fixed sample size $n = 2$. At the first step, a vector $\mathbf{u}(0)$ must be chosen. This vector may be chosen freely but must be such that $\pi + \mathbf{u}(0)$ remains in the subspace of constraints. Actually, the cube method is a family of methods that depends on the way the vector $\mathbf{u}(0)$ is chosen. This vector may be chosen randomly or not.

If, from π, we follow the direction given by vector $\mathbf{u}(0)$, then we will necessarily cross a face of the cube. Let us consider this point denoted on Figure 4 by $\pi(0) + \lambda_1^*(0)\mathbf{u}(0)$. Now, if, from π, we follow the opposite direction, *i.e.*, the direction given by vector $-\mathbf{u}(0)$, we will also cross a face of the cube. Let us consider this point denoted on Figure 4 by $\pi(0) - \lambda_2^*(0)\mathbf{u}(0)$. At the first step, vector $\pi(0) = \pi$ is modified randomly. Vector $\pi(1)$ will be set to $\pi(0) + \lambda_1^*(0)\mathbf{u}(0)$ or to $\pi(0) - \lambda_2^*(0)\mathbf{u}(0)$. The choice is done randomly in such a way that $\mathrm{E}[\pi(1)] = \pi(0)$. At the end of the first step of the flight phase, we have thus jumped on a face of the cube, which means that at least one component of $\pi(1)$ is equal to 0 or 1, *i.e.*, the problem is reduced from a problem of sampling from a population of size $N = 3$ to a population of size $N = 2$. In N steps at least, the flight phase is thus completed.

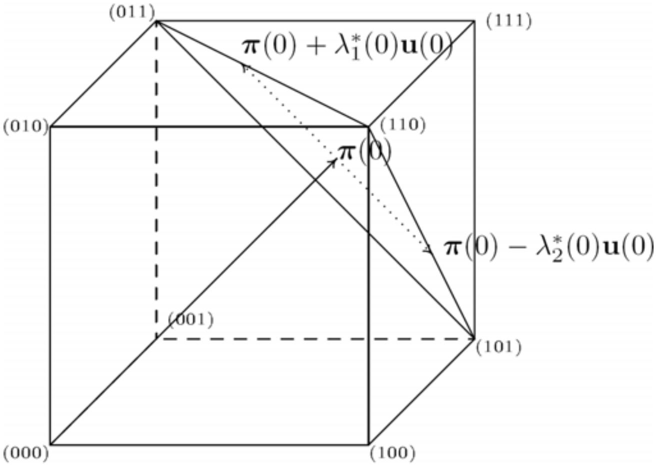

(011)
(111)
$\pi(0) + \lambda_1^*(0)\mathbf{u}(0)$
(010)
(110)
$\pi(0)$
$\pi(0) - \lambda_2^*(0)\mathbf{u}(0)$
(001)
(101)
(000)
(100)

Figure 4 Flight phase in a population of size $N = 3$ with a sample size constraint $n = 2$

More generally, the flight phase is a random walk in the intersection of the balancing subspace and the cube. This random walk stops at a vertex of the intersection of the cube and the subspace. The flight phase is defined by the following class of algorithms. First initialize with $\pi(0) = \pi$. Next, at time $t = 0, ..., T$,

1. Generate any vector $\mathbf{u}(t) = [u_k(t)] \neq 0$ such that
   (i) $\mathbf{u}(t)$ is in the kernel of matrix $\mathbf{A} = (\mathbf{x}_1/\pi_1, \ldots, \mathbf{x}_k/\pi_k, \ldots, \mathbf{x}_N/\pi_N)$, *i.e.*, $\mathbf{Au}(t) = 0$,
   (ii) $u_k(t) = 0$ if $\pi_k(t)$ is integer.
2. Compute $\lambda_1^*(t)$ and $\lambda_2^*(t)$, the largest values such that
$$0 \leq \pi(t) + \lambda_1(t)\mathbf{u}(t) \leq 1,$$
$$0 \leq \pi(t) - \lambda_2(t)\mathbf{u}(t) \leq 1.$$
3. Compute
$$\pi(t+1) = \begin{cases} \pi(t) + \lambda_1^*(t)\mathbf{u}(t) & \text{with probability } q_1(t) \\ \pi(t) - \lambda_2^*(t)\mathbf{u}(t) & \text{with probability } q_2(t), \end{cases}$$
where $q_1(t) = \lambda_2^*(t)/\{\lambda_1^*(t) + \lambda_2^*(t)\}$ and $q_2(t) = 1 - q_1(t)$.

The flight phase stops when it is no longer possible to find a vector $\mathbf{u}(t) \neq 0$.

5.3 Landing phase

If, at the end of the flight phase, the balancing equations are not exactly satisfied, there is a need for a landing phase. Let $\pi^* = [\pi_k^*]$ be the vector obtained at the last step of the flight phase. It is possible to prove (see Deville and Tillé 2004) that

$$\mathrm{card}(U^*) \leq p,$$

where

$$U^* = \{k \in U \mid 0 < \pi_k^* < 1\}$$

and p is the number of balancing variables. The aim of the landing phase is to find a sample $\mathbf{s}$ such that

$\mathrm{E}(\mathbf{s}|\boldsymbol{\pi}^*) = \boldsymbol{\pi}^*$, which is almost balanced. There are two ways of selecting such a sample:

1. *The flight phase by linear programming* consists of considering all the possible samples of U^*. A cost is assigned to each sample. This cost, is, for instance, the distance between the sample and the subspace of constraints. Next, one looks for a sampling design on U^* that minimizes the expected cost and that satisfies the inclusion probabilities $\boldsymbol{\pi}^*$. This problem can be solved because the number of samples to consider is reasonable due to the small size of U^*.
2. *The flight phase by suppression of variables* may be used when the number of balancing variables is too large for the linear program to be solved by a simplex algorithm $(p > 20)$. With this method, an auxiliary variable is dropped at the end of the flight phase. Next, we can return to the flight phase until it is no longer possible to 'move' within the constraint subspace. The constraints are then relaxed successively according to an order of preference.

6. Variance and variance estimation

6.1 A residual technique

The variance of the Horvitz-Thompson estimator can be estimated by using a residual technique developed in Deville and Tillé (2005). The residual technique is comparable to the technique used to estimate the variance of the calibration estimator and has been validated by a set of simulations. The estimated variance of the Horvitz-Thompson estimator is thus very similar to the estimated variance of a generalized regression (GREG) estimator. Nevertheless, the variance of the GREG estimator is generally underestimated because it does not take into account the randomness of the weights. Indeed, if the usual variance of the GREG estimator is computed for the special case of poststratification, we obtain the variance of a stratified design with proportional allocation. The variance of the poststratified estimator is nevertheless larger than the variance in a stratified design with proportional allocation.

6.2 Approximation of variance

If the balanced sampling design has a large entropy, Hájek (1981) and Deville and Tillé (2005, method 4) have proposed the following approximation of the design variance given by:

$$\mathrm{var}_p(\widehat{Y}_\pi) \cong \mathrm{var}_{app}(\widehat{Y}_\pi) = \sum_{k\in U} d_k \frac{(y_k - \mathbf{x}_k'\mathbf{b})^2}{\pi_k^2}, \quad (5)$$

where the subscript p denotes the sampling design,

$$\mathbf{b} = \left(\sum_{k\in U} d_k \frac{\mathbf{x}_k\mathbf{x}_k'}{\pi_k^2}\right)^{-1} \sum_{k\in U} d_k \frac{\mathbf{x}_k y_k}{\pi_k^2},$$

and the d_k are the solution of the nonlinear system

$$\pi_k(1-\pi_k) = d_k - \frac{d_k\mathbf{x}_k'}{\pi_k}\left(\sum_{\ell\in U} d_\ell \frac{\mathbf{x}_\ell\mathbf{x}_\ell'}{\pi_\ell^2}\right)^{-1} \frac{d_k\mathbf{x}_k}{\pi_k}, \; k \in U. \quad (6)$$

The entropy of the sampling design depends on the way vectors $\mathbf{u}(t)$ are chosen during the flight phase. In order to increase the entropy, vector $\mathbf{u}(t)$ can be chosen randomly or the population can be randomly sorted before selecting the sample.

Expression (5), which only uses the first-order inclusion probabilities, was validated by Deville and Tillé (2005) under a variety of balanced samples regardless of how the *y*-values were generated. An approximation very close to Expression (5) was obtained by Fuller (2009) and Legg and Yu (2010) for a balanced sampling design obtained by a rejective procedure in the case of an initial design that uses Poisson sampling. These approximations do not take the rounding problem into account.

6.3 Estimation of variance

Deville and Tillé (2005) proposed a family of variance estimators for balanced sampling, of the form

$$\widehat{\mathrm{var}}(\widehat{Y}_\pi) = \sum_{k\in S} c_k \frac{(y_k - \mathbf{x}_k'\widehat{\mathbf{b}})^2}{\pi_k^2}, \quad (7)$$

where

$$\widehat{\mathbf{b}} = \left(\sum_{\ell\in S} c_\ell \frac{\mathbf{x}_\ell\mathbf{x}_\ell'}{\pi_\ell^2}\right)^{-1} \sum_{\ell\in S} c_\ell \frac{\mathbf{x}_\ell y_\ell}{\pi_\ell^2}$$

and the c_k are the solutions of the nonlinear system

$$1 - \pi_k = c_k - \frac{c_k\mathbf{x}_k'}{\pi_k}\left(\sum_{\ell\in S} c_\ell \frac{\mathbf{x}_\ell\mathbf{x}_\ell'}{\pi_\ell^2}\right)^{-1} \frac{c_k\mathbf{x}_k}{\pi_k}, \quad (8)$$

which can be solved by a fixed point algorithm.

In Deville and Tillé (2005), simpler variants of c_k were also proposed. For instance, one can use the alternative values,

$$\tilde{c}_k \approx \frac{n}{n-p}(1-\pi_k),$$

that are very close to c_k. The estimator $\widehat{\mathrm{var}}(\widehat{Y}_\pi)$ is approximately design-unbiased because it is an estimator by substitution of the approximation given in expression (5), (for more information regarding estimators obtained by substitution, see Deville 1999), which is a reasonable approximation of the variance under the sampling design.

It is not easy to use bootstrap method to estimate the variance in the context of balanced sampling. Balanced samples with replacement should be selected from the original sample. A generalization of the cube method for balanced sampling with replacement has not yet been described. A solution, proposed by Chauvet (2007), consists of reconstructing an artificial population from the sample.

Next, bootstrap samples are selected by using balanced sampling. Another solution was proposed by Fuller (2010) for balanced rejective sampling. Breidt and Chauvet (2010a) have proposed an alternative method where a martingale difference representation of the cube method is used in order to approximate second-order inclusion probabilities, which enables us to construct a nearly unbiased variance estimator.

7. Balanced sampling in practice

7.1 Interest of balanced sampling

In the model-assisted and the model-based frameworks, a balancing sampling design with the Horvitz-Thompson estimator is often the optimal strategy (see Nedyalkova and Tillé 2009). Indeed, when the sample is balanced, the variances of the Horvitz-Thompson estimators of the auxiliary variables are equal to zero. Under a linear model, the variance of the Horvitz-Thompson estimator of the interest variable will only depend on the residuals of the model.

The advantages of balanced sampling are as follows:

(i) Balanced sampling increases the accuracy of the Horvitz-Thompson estimator. This point has been developed in Section 6. Indeed, the variance of the Horvitz-Thompson estimator only depends on the residuals of the regression of the interest variable by the balancing variables.

(ii) Balanced sampling protects against large sampling errors. Indeed, the most unfavourable samples have a null probability of being selected.

(iii) If the variable of interest is well explained by the auxiliary information, in model-based inference, balanced sampling protects against a mis-specification of the model. This point is largely developed by Royall (1976b, a) and Valliant *et al.* (2000). A recent discussion of this important question is given in Nedyalkova and Tillé (2009, 2010).

(iv) Balanced sampling can ensure that the sample sizes in planned domains are not too small or - much worse - equal to zero. Indeed, if an indicator variable of the domain is added in the list of auxiliary variables, the size of the domain is then fixed in the sample.

(v) Balanced sampling allows us to avoid random weights. With balanced sampling, the Horvitz-Thompson weights can be used. If the sampling design does not contain any balancing constraints (for instance with Poisson sampling) the weighting system obtained by a calibration procedure becomes very random, which increases the variance of the estimators. If the sample is balanced, the weights will be less random even if a calibration procedure is used after balancing.

The availability of easy to use packages contributed to the large use of the cube method in several important statistical processes. The first main application of the cube method is selection of the rotation groups for the French census. (See Desplanques 2000; Dumais, Bertrand and Kauffmann 2000; Durr and Dumais 2001, 2002; Dumais and Isnard 2000; Bertrand, Christian, Chauvet and Grosbras 2004; da Silva, da Silva Borges, Aires Leme and Moura Reis Miceli 2006). For the municipalities with fewer than 10,000 inhabitants, five non-overlapping rotation groups of municipalities are selected using a balanced sampling design with equal inclusion probabilities (1/5). Each year, a fifth of the municipalities are surveyed. So after 5 years, all the small municipalities are selected. For the municipalities with more than 10,000 inhabitants, in each municipality, five non-overlapping balanced samples of addresses are selected with inclusion probabilities 8%. So, after 5 years, 40% of the addresses are visited. The balancing variables are socio-demographic variables taken from the last census.

In the French master sample, the primary units are geographical areas that are selected using a balanced sampling design (see Wilms 2000; Christine and Wilms 2003; Christine 2006). The master sample is a self-weighted multi-stage sampling. So the primary units are selected with unequal probabilities that are proportional to their sizes. The balancing variables are socio-demographic variables taken from the last census. Bardaji (2001) and Even (2002) have also used balanced sampling to select a sample of beneficiaries of subsidized jobs. Seven populations are surveyed, a balanced sample of beneficiaries is selected in each of the populations by using between two and five balancing variables according to the populations.

In the company Électricité de France (EDF), new electricity meters allow electricity consumption for each household to be measured on a continuous basis. The amount of information collected is so large that it is impossible to archive all the data. Dessertaine (2006, 2007) used balanced sampling to select the time series of consumption that must be archived in order to ensure that they represent the consumption of the entire French population as accurately as possible. Biggeri and Falorsi (2006) used balanced sampling to improve the quality of the consumer price index in Italy. Gismondi (2007) tested balanced sampling to estimate the number of tourist nights spent in Italy. D'Alò, Di Consiglio, Falorsi and Solari (2006) and Falorsi and Righi (2008) also proposed using a balanced sampling design to estimate totals in small domains. Simulations were run by Marí, Barbará, Mitas and Passamonti (2007b, a) in Argentina and Chipperfield (2009) in Australia to assess the interest of balanced sampling for the master sample.

At Statistics Canada, Fecteau and Jocelyn (2006) and Jocelyn (2006) tested balanced sampling to select a sample of businesses. Canadian unincorporated businesses complete their income tax returns either on paper or electronically. More than half of the returns are submitted electronically. Balanced sampling was used to select a sample from businesses that responded electronically so that, for some key variables that are known for the whole population, the sample means matched the known population means.

Balanced sampling can also be used to impute a missing value in case of item nonresponse. Indeed, using a model to predict an imputation allocates central values, which will lead to a biased inference on quantiles. In contrast, a random imputation generally increases the variances of the estimators. In order to solve this dilemma, Deville (1998, 2005, 2006) and Chauvet, Deville and Haziza (2010c, b) have proposed using imputation by prediction and to add a residual that is chosen amongst the residuals of the respondent according to a balanced sampling design. This is done to avoid adding a term of variance to the total of the imputed variable.

7.2 Balanced sampling versus other sampling techniques

Unequal probability sampling is a particular case of the cube method. Indeed, when the only auxiliary variable is the inclusion probability, the sample has a fixed sample size. The cube method is a generalization of the splitting method (see Deville and Tillé 1998), which includes several sampling algorithms with unequal probabilities (Brewer's method, pivotal method, corrected Sunter method, see Brewer 1975; Sunter 1977; Deville and Tillé 1998; Tillé 2006b). Stratification is also a particular case of balanced sampling. With the cube method, one can balance on overlapping strata and use qualitative and quantitative variables together. Systematic sampling can even be seen as a balanced sampling design on the order statistic related to the variable on which the population is ordered.

Almost all the other sampling techniques are particular cases of balanced sampling (except multistage sampling). In fact, balanced sampling is simply more general, in the sense that all the other methods of sampling can be implemented with the cube method. The cube method allows us to use any variable for balancing with a reasonable computation time. With the more general concept of balancing, strata can overlap, quantitative and qualitative variables can be used together, and the inclusion probabilities can be chosen freely.

It is well known that the ratio estimator and the post-stratified estimator are particular cases of the regression estimator. The regression estimator is also a particular case of the calibration estimator (which includes a non-linear adjustment). In the same way, balanced sampling is a more general method of sampling that includes almost all the other methods. The algorithm of the cube method may seem complicated but, once implemented, it enables us to run a function with two arguments: the vector of inclusion probabilities and the matrix of balancing variables.

7.3 Choice of the balancing strategy

The main recommendation is to choose balancing variables that are closely correlated to the interest variables. As with any regression problem, the balancing variables must be chosen parsimoniously: one must not choose too many balancing variables because, accuracy no longer improves with a large number of variables and the instability of the variance estimator increases with each additional variable. Practically, the aim is not to estimate one variable but a set of interest variables. Thus, the set of auxiliary variables must be correlated to all the interest variables. Moreover, the auxiliary variables should not be too correlated amongst themselves.

Lesage (2008) has proposed a method to balance a sample on complex statistics rather than simply using population totals. The main idea consists in balancing on the linearized value (or influence function) of the parameter of interest. Breidt and Chauvet (2010b) have proposed using penalized balanced sampling in order to possibly relax some balancing constraints, which can be useful for instance in small domain estimation.

In many cases, the balancing variables contain measurement errors. For example, in most registers, one can suspect errors in the data. Missing values can obviously occur and auxiliary variables are often corrected by a method of imputation. As for calibration, the fact of having errors in the auxiliary variables is not very important as long as the calibration is done on the total of the auxiliary variables of the register. Indeed, with balanced sampling, the Horvitz-Thompson estimator is used and is unbiased even if the auxiliary variables are false. The gain in efficiency only depends on the correlation between the balancing variables and the interest variables. This correlation is rarely affected by errors in the balancing variables.

Several variables can be used to improve small domain estimates. To ensure that a domain D is not empty, one can simply add the auxiliary variable:

$$x_k = \begin{cases} \pi_k & \text{if } k \in D \\ 0 & \text{otherwise,} \end{cases}$$

which implies that the number of sampled units that belong to D is equal to

$$n_D = \sum_{k \in U} x_k = \sum_{k \in D} \pi_k,$$

if n_D is integer, or one of the closest two integers to n_D if n_D is not an integer.

In some cases, it is interesting to balance on auxiliary variables in subgroups, domains or strata. An interesting procedure described in Chauvet (2009) consists of separately running the flight phase in each stratum. A rounding problem will then occur in each stratum. These rounding problems can then be merged and a flight phase can be run again on the whole population. Finally, the landing phase is applied only to the whole population. This procedure enables us to roughly satisfy the balancing equations in each strata without cumulating the rounding problems.

The inclusion probabilities must be computed prior to sampling. When a linear model is assumed, these probabilities should in principle be proportional to the errors of the model in order to minimize variance (see Tillé and Favre 2005; Chauvet, Bonnery and Deville 2010a; Nedyalkova and Tillé 2009, 2010). This choice generalizes Neyman's allocation for stratified sampling (Neyman 1934). However, the inclusion probabilities often need to be chosen on others constraints. For instance, in order to construct the rotation groups of the French census, the inclusion probabilities must all be equal to a fifth.

7.4 Balancing versus calibration

Stratification is a particular case of balancing, while post-stratification is a particular case of calibration. In stratification and balancing, the weights do not become random. It is thus generally a better strategy. Nevertheless, more auxiliary information is needed for balancing. Indeed, for balanced sampling, the auxiliary variables must be known for all the units of the population, whereas, for calibration, only the population totals are needed. Balancing is a very interesting method for small population sizes. It is thus a very good method for selecting primary units in a multistage sampling design.

Both techniques can be used together. They are not contradictory. The best strategy consists of using balanced sampling and calibration together. Indeed calibration can resolve the small rounding problem that may remain after balancing. At the estimation stage, more auxiliary variables are often available because, in order to balance a sample, the auxiliary information must be known at the individual level while, in order to calibrate the sample, only the population totals are necessary.

Generally, it is recommended to re-calibrate on the balancing variables at the estimation stage even if more calibration variables are available. If only new variables are used in calibration, the effect of balancing can be lost. There is, however, one case where calibration can be used without re-calibrating on the balancing variables: when, conditionally on the calibration variables, we can reasonably assume that the balancing variables are no longer correlated to the variables of interest. This can occur when the balancing and the calibration variables are the same variables measured at different moments, and the calibration variables are more recent.

When the determination coefficient between the interest variable and the auxiliary variables is equal to or close to one, then calibration is more efficient because of the rounding problem of balanced sampling. Anyway the most efficient strategy always consists of using balanced sampling and calibration together (see the simulation in Deville and Tillé 2004).

7.5 Accuracy of the balancing equations

It is possible to prove, under realistic assumptions (see Deville and Tillé 2004), that with the cube method

$$\left|\frac{\widehat{X}_j - X_j}{X_j}\right| < O(p/n),$$

where p is the number of variables, and $O(x)/x$ is a quantity that remains bounded when x tends to infinity. With simple random sampling

$$\left|\frac{\widehat{X}_j - X_j}{X_j}\right| = O_p(\sqrt{1/n}),$$

where $O_p(x)/x$ is a quantity that remains bounded in probability when x tends to infinity.

The gains in accuracy are therefore considerable. The small rounding problem can be fixed by a small calibration. The rounding problem comes from the fact that selecting a sample is an integer problem. It also occurs in stratification, which is a particular case of balancing. In stratification with proportional allocation, the sums of the inclusion probabilities in the strata are generally not integers. So, the sample sizes in the strata are obtained by rounding the sum of inclusion probabilities in the strata. The cube method does this rounding automatically and randomly in such a way as to ensure that the inclusion probabilities are exactly satisfied.

7.6 Balanced sampling in repeated surveys

An important difficulty occurs in repeated sampling. The problem comes from the fact that, when a balanced sample is selected with unequal inclusion probabilities, the complementary sample is not necessarily balanced. Indeed, the equality

$$\sum_{k\in S}\frac{\mathbf{x}_k}{\pi_k} = \sum_{k\in U}\mathbf{x}_k$$

does not imply that

$$\sum_{k\in U\setminus S}\frac{\mathbf{x}_k}{1-\pi_k} = \sum_{k\in U}\mathbf{x}_k.$$

This problem occurred in the French master sample. In this sampling design, the primary units, which are geographical areas, are selected with unequal probabilities that are proportional to the size. After selecting the sample, some regions asked for complementary samples of areas that were not already selected. This question is intricate, because the complementary sample of a balanced sample is no longer balanced, and the aim is thus to select a balanced sample from a part of the population that is no longer balanced. Tillé and Favre (2004) gave a few methods to co-ordinate balanced samples, which were selected with unequal inclusion probabilities. More generally, the coordination (in the sense of managing overlap) of balanced samples can be difficult when the sampling design is balanced.

While challenging, it is possible to organize rotations if all the samples are selected together and the samples are selected with equal inclusion probabilities. Indeed, in this case the complementary $\overline{S} = U \setminus S$ of the samples S is also a balanced sample. A second balanced sample can be directly selected from $\overline{S}$ and so on. This method was used to create five rotation groups in the French master sample. The five groups are five balanced samples of municipalities.

If the samples are selected with unequal inclusion probabilities, some solutions are described in Tillé and Favre (2004). An interesting particular case can easily be solved: when two non-overlapping samples must be selected with the same unequal inclusion probabilities $\pi_k < 0.5$ from the same population. First, a sample S_A balanced on $\mathbf{x}_k$ must be selected with inclusion probabilities $\pi_{kA} = 2\pi_k$ such that

$$\sum_{k\in S_A} \frac{\mathbf{x}_k}{2\pi_k} = \sum_{k\in U} \mathbf{x}_k.$$

Next, a sample S_1 can be selected from S_A. This sample must be selected with inclusion probability $\pi_{kB} = 0.5$ and must be balanced on $\mathbf{x}_k/2\pi_k$, which gives the following balancing equations:

$$\sum_{k\in S_2} \frac{\mathbf{x}_k/(2\pi_k)}{1/2} = \sum_{k\in S_A} \frac{\mathbf{x}_k}{2\pi_k} = \sum_{k\in U} \mathbf{x}_k.$$

The sample $S_2 = S_A \setminus S_1$ is also balanced.

If the population changes over times (deaths and births), the organization of a rotation becomes much more difficult. This difficulty already occurs with stratified samples. Nevertheless, for stratification, several reasonable solutions exist (see, amongst others, De Ree 1999; Hesse 1998; Rivière 1999; Nedyalkova, Péa and Tillé 2006).

7.7 Main implementations of balanced sampling

An SAS/IML® implementation was first programmed by three students of the École nationale de la statistique et de l'analyse de l'information (Ensai) (Bousabaa *et al.* 1999). An official version of the *Institut National de la Statistique et des Études Économiques* done by Tardieu (2001) and Rousseau and Tardieu (2004) is now available on the Insee Web site. Another SAS/IML® version done by Chauvet and Tillé (2005b, a, 2006) is also available on the University of Neuchâtel Web site. In R language, the sampling package (Tillé and Matei 2007) allows us to use the cube method. These software programs are free, available over the Internet and are easy to use.

The available programs written using R language or SAS/IML® have no limit as far as population size is concerned. An application with 40 balanced variables is possible. In order to select the sample, the computation times increase with $N \times p^2$, where N is the population size and p the number of balancing variables. It is thus possible to select a sample in a population of several million statistical units.

Acknowledgements

This paper has been written in response to an invitation to speak at the Demographic Statistical Methods Division Seminar of the U.S. Census Bureau in June 2008. The author would like to thank the U.S. Census Bureau and particularly Patrick Flanagan without whom this paper would never have been written. The author is also grateful to an associate editor and two anonymous reviewers for valuable comments and corrections that helped to improve this paper.

References

Ardilly, P. (1991). Échantillonnage représentatif optimum à probabilités inégales. *Annales d'Économie et de Statistique*, 23, 91-113.

Ardilly, P. (2006). *Les Techniques de Sondage*. Technip, Paris.

Bardaji, J. (2001). Un an après la sortie d'un contrat emploi consolidé : près de six chances sur dix d'avoir un emploi. *Premières Informations Synthèses*, *Direction de l'Animation de la Recherche des Etudes et des Statistiques* (*DARES*) *du Ministère du Travail des relations sociales et de la solidarité*, 43, 3, 1-8.

Bertrand, P., Christian, B., Chauvet, G. and Grosbras, J.-M. (2004). Plans de sondage pour le recensement rénové de la population. In *Séries Insee Méthodes* : *Actes des Journées de Méthodologie Statistique*, Paris. Insee.

Biggeri, L., and Falorsi, P.D. (2006). A probability sample strategy for improving the quality of the consumer price index survey using the information of the business register. In *Proceedings of the Conference of European Statisticians Group of Experts on Consumer Price Indices*, Eighth Meeting, Geneva, 10-12 May 2006.

Bousabaa, A., Lieber, J. and Sirolli, R. (1999). La macro cube. Technical report, Ensai, Rennes.

Breidt, F.J., and Chauvet, G. (2010a). Improved variance estimation for balanced samples drawn via the cube method. *Journal of Statistical Planning and Inference*, 141, 479-487.

Breidt, F.J., and Chauvet, G. (2010b). Penalized balanced sampling. Working paper, Ensai.

Brewer, K.R.W. (1975). A simple procedure for πpswor. *Australian Journal of Statistics*, 17, 166-172.

Brewer, K.R.W. (1999). Design-based or prediction-based inference? Stratified random *vs* stratified balanced sampling. *International Statistical Review*, 67, 35-47.

Chauvet, G. (2007). *Méthodes de Bootstrap en Population Finie*. PhD thesis, Université Rennes 2.

Chauvet, G. (2009). Stratified balanced sampling. *Survey Methodology*, 35, 115-119.

Chauvet, G., Bonnery, D. and Deville, J.-C. (2010a). Optimal inclusion probabilities for balanced sampling. *Journal of Statistical Planning and Inference*, 141, 2, 984-994.

Chauvet, G., Deville, J. and Haziza, D. (2010b). Adapting the cube algorithm for balanced random imputation in surveys. Technical report, Ensai, Rennes.

Chauvet, G., Deville, J. and Haziza, D. (2011). On balanced random imputation in surveys. *Biometrika*.

Chauvet, G., and Tillé, Y. (2005a). *Fast SAS Macros for balancing Samples*: *user's guide*. Software Manual, University of Neuchâtel, http://www2.unine.ch/statistics/page10890.html.

Chauvet, G., and Tillé, Y. (2005b). New SAS macros for balanced sampling. In *Journées de Méthodologie Statistique*, Insee, Paris.

Chauvet, G., and Tillé, Y. (2006). A fast algorithm of balanced sampling. *Journal of Computational Statistics*, 21, 9-31.

Chipperfield, J. (2009). An evaluation of cube sampling for ABS household surveys. Technical report, Australian Bureau of Statistics.

Christine, M. (2006). Use of balanced sampling in the framework of the master sample for french household surveys. In *Joint Statistical Meeting of the American Statistical Association*, Seattle August 2006.

Christine, M., and Wilms, L. (2003). Theoretical and practical problems related to the development of "EMEX": How to improve the precision of the regional supplements of National Surveys with an Additional Sample? In *Proceedings*: *Symposium 2003*, *Challenges in Survey Taking for the Next Decade*, Statistics Canada, Ottawa.

Cumberland, W.G., and Royall, R.M. (1981). Prediction models in unequal probability sampling. *Journal of the Royal Statistical Society*, B, 43, 353-367.

Cumberland, W.G., and Royall, R.M. (1988). Does simple random sampling provide adequate balance? *Journal of the Royal Statistical Society*, B, 50, 118-124.

da Silva, A.D., da Silva Borges, A., Aires Leme, R. and Moura Reis Miceli, A.P. (2006). Modalidades alternativas de censo demográfico: o cenário internacional a partir das experiências dos estados unidos, frança, holanda, israel e alemanha. Technical report, Instituto Brasileiro de Geografia e Estatística.

D'Alò, M., Di Consiglio, L., Falorsi, S. and Solari, F. (2006). Small area estimation of the italian poverty rate. *Statistics in Transition*, 7, 771-784.

De Ree, S.J.M. (1999). Co-ordination of business samples using measured response burden. In *Contributed paper*, *52th Session of the ISI Helsinki*.

Desplanques, G. (2000). La rénovation du recensement de la population. In *Actes de la séance du 5 octobre 2000 du séminaire méthodologique SFDS-Insee sur la rénovation du recensement*, 2-5.

Dessertaine, A. (2006). Sondages et séries temporelles : une application pour la prévision de la consommation électrique. In *Actes des journées Françaises de Statistique 2006*, Clamart, France.

Dessertaine, A. (2007). Sampling and data-stream: Some ideas to built balanced sampling using auxiliary Hilbertian informations. In *Proceedings of 56th the International Statistical Institute Conference*: *IPM56 - New methods of sampling*, Lisboa, Portugal.

Deville, J.-C. (1992). Constrained samples, conditional inference, weighting: Three aspects of the utilisation of auxiliary information. In *Proceedings of the Workshop on the Uses of Auxiliary Information in Surveys*, Örebro (Sweden).

Deville, J.-C. (1998). La correction de la non-réponse par calage ou par échantillonnage équilibré. In *Recueil de la Section des méthodes d'enquêtes des communications présentées au 26ème congrès de la Société Statistique du Canada*, 103-110, Sherbrooke.

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

Deville, J.-C. (2005). Imputation stochastique et échantillonnage équilibré. Technical report, École Nationale de la Statistique et de l'Analyse de l'Information.

Deville, J.-C. (2006). Stochastic imputation using balanced sampling. In *Joint Statistical Meeting of the American Statistical Association*, Seattle August 2006.

Deville, J.-C., Grosbras, J.-M. and Roth, N. (1988). Efficient sampling algorithms and balanced sample. In *COMPSTAT*, *Proceedings in Computational Statistics*, Heidelberg. Physica Verlag, 255-266.

Deville, J.-C., and Tillé, Y. (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika*, 85, 89-101.

Deville, J.-C., and Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, 91, 893-912.

Deville, J.-C., and Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 128, 569-591.

Dudoignon, L., and Vanheuverzwyn, A. (2006). Tirage d'un échantillon à probabilités inégales : application au panel Médiamat. In *Actes de des Journées de Méthodologie Statistique*, 1-10.

Dumais, J., Bertrand, P. and Kauffmann, B. (2000). Sondage, estimation et précision dans la rénovation du recensement de la population. In *Actes de la séance du 5 octobre 2000 du séminaire méthodologique SFDS-Insee sur la rénovation du recensement*, 6-26.

Dumais, J., and Isnard, M. (2000). Le sondage de logements dans les grandes communes dans le cadre du recensement rénové de la population. In *Séries Insee Méthodes* : *Actes des Journées de Méthodologie Statistique*, Paris. Insee, 100, 37-76.

Durr, J.-M., and Dumais, J. (2001). Redesign of the french census of population. In *Proceedings*: *Symposium 2001*, *Achieving Data Quality in a Statistical Agency*: *A Methodological Perspective*, Statistics Canada, Ottawa.

Durr, J.-M., and Dumais, J. (2002). Redesign of the french census of population. *Survey Methodology*, 28, 43-49.

Even, K. (2002). Improved tool for evaluating employment and vocational training policy: Panel of beneficiaries. *Premières Informations Synthèses*, *Direction de l'Animation de la Recherche des Études et des Statistiques (DARES) du Ministère du Travail des relations sociales et de la solidarité*, 33, 1, 1-7.

Falorsi, P.D., and Righi, P. (2008). A balanced sampling approach for multi-way stratification designs for small area estimation. *Survey Methodology*, 34, 223-234.

Fecteau, S., and Jocelyn, W. (2006). Une application de l'échantillonnage équilibré : le plan de sondage des entreprises non incoprporées. In *Méthodes d'enquêtes et sondages* : *pratiques européenne et nord-américaine*, (Eds., P. Lavallée and L.-P. Rivest), Paris. Dunod, 405-410.

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

Fuller, W.A. (2010). Replication variance estimation for rejective sampling. In *Seminar of Statistics Canada*, June 2010, Ottawa.

Gini, C. (1928). Une application de la méthode représentative aux matériaux du dernier recensement de la population italienne (1er décembre 1921). *Bulletin of the International Statistical Institute*, 23, 2, 198-215.

Gini, C., and Galvani, L. (1929). Di una applicazione del metodo rappresentative all'ultimo censimento Italiano della popolazione (1° decembri, 1921). *Annali di Statistica*, Series 6, 4, 1-107.

Gismondi, R. (2007). Quick estimation of tourist nights spent in italy. *Statistical Methods and Applications*, 16, 141-168.

Hájek, J. (1981). *Sampling from a Finite Population*. New York: Marcel Dekker.

Hedayat, A.S., and Majumdar, D. (1995). Generating desirable sampling plans by the technique of trade-off in experimental design. *Journal of Statistical Planning and Inference*, 44, 237-247.

Hesse, C. (1998). Sampling co-ordination: A review by country. Technical Report E9908, Direction des Statistique d'Entreprises, Insee, Paris.

Jocelyn, W. (2006). Sampling and estimation strategies for the canadian unincorporated business population. In *Joint Statistical Meeting of the American Statistical Association*, Seattle August 2006.

Kiaer, A. (1896). Observations et expériences concernant des dénombrements représentatifs. *Bulletin de l'Institut International de Statistique*, 9, 2, 176-183.

Kiaer, A. (1899). Sur les méthodes représentatives ou typologiques appliquées à la statistique. *Bulletin de l'Institut International de Statistique*, 11, 1, 180-185.

Kiaer, A. (1903). Sur les méthodes représentatives ou typologiques appliquées à la statistique. *Bulletin de l'Institut International de Statistique*, 13, 1, 66-78.

Kiaer, A. (1905). Discours sans intitulé sur la méthode représentative. *Bulletin de l'Institut International de Statistique*, 14, 1, 119-134.

Kott, P.S. (1986). When a mean-of-ratios is the best linear unbiased estimator under a model. *The American Statistician*, 40, 202-204.

Langel, M., and Tillé, Y. (2010). Corrado Gini, a pioneer in balanced sampling and inequality theory. Technical report, University of Neuchatel.

Legg, J.C., and Yu, C.L. (2010). A comparison of sample set restriction procedures. *Survey Methodology*, 36, 69-79.

Lesage, E. (2008). Contraintes d'équilibrage non linéraires. In *Méthodes d'enquêtes : applications aux enquêtes longitudinales, à la santé et aux enquêtes électorales*, (Eds., P. Guilbert, D. Haziza, A. Ruiz-Gazen and Y. Tillé), Paris. Dunod, 285-289.

Marí, G., Barbará, G., Mitas, G. and Passamonti, S. (2007a). Construcción de un estimador de variancia para muestras balanceadas estratificadas. In *XXXV Coloquio Argentino de Estadística. Mar del Plata*, Argentina. 22, 23 y 24 de Octubre de 2007.

Marí, G., Barbará, G., Mitas, G. and Passamonti, S. (2007b). Muestras equilibradas en poblaciones finitas: un estudio comparativo en muestras de explotaciones agropecuarias. In *Undécimas Jornadas "Investigaciones en la Facultad" de Ciencias Económicas y Estadística, noviembre de 2007*, Universidad Nacional de Rosario, Argentina.

Nedyalkova, D., Péa, J. and Tillé, Y. (2006). A review of some current methods of coordination of stratified samples. introduction and comparison of new methods based on microstrata. Technical report, Université de Neuchâtel.

Nedyalkova, D., and Tillé, Y. (2009). Optimal sampling and estimation strategies under linear model. *Biometrika*, 95, 521-537.

Nedyalkova, D., and Tillé, Y. (2010). Bias robustness and efficiency in model-based inference. Technical report, University of Neuchâtel.

Neyman, J. (1934). On the two different aspects of representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-606.

Neyman, J. (1952). *Lectures and Conferences on Mathematical Statistics and Probability*. Graduate School; U.S. Department of Agriculture, Washington.

Périé, P. (2008). Échantillonnage à entropie maximale sous contraintes : un algorithme rapide basé sur l'optimisation linéaire en nombres binaires. In *Méthodes d'enquêtes : applications aux enquêtes longitudinales, à la santé et aux enquêtes électorales*, (Eds., P. Guilbert, D. Haziza, A. Ruiz-Gazen and Y. Tillé), Paris. Dunod, 294-299.

Rivière, P. (1999). Coordination of samples: The microstrata methodology. In *13th International Roundtable on Business Survey Frames*, Paris. Insee.

Rousseau, S., and Tardieu, F. (2004). La macro SAS CUBE d'échantillonnage équilibré, Documentation de l'utilisateur. Technical report, Insee, Paris.

Royall, R.M. (1976a). Likelihood functions in finite population sampling theory. *Biometrika*, 63, 605-614.

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

Royall, R.M. (1988). The prediction approach to sampling theory. In *Handbook of Statistics Volume 6*: *Sampling*, (Eds., P.R. Krishnaiah and C.R. Rao), Amsterdam. Elsevier/North-Holland, 399-413.

Royall, R.M., and Pfeffermann, D. (1982). Balanced samples and robust bayesian inference in finite population sampling. *Biometrika*, 69, 401-409.

Sunter, A. (1977). List sequential sampling with equal or unequal probabilities without replacement. *Applied Statistics*, 26, 261-268.

Tardieu, F. (2001). Échantillonnage équilibré: de la théorie à la pratique. Technical report, Insee, Paris.

Thionet, P. (1953). *La théorie des sondages*. Insee, Imprimerie nationale, Paris.

Tillé, Y. (2001). *Théorie des sondages* : *échantillonnage et estimation en populations finies*. Dunod, Paris.

Tillé, Y. (2006a). Balanced sampling by means of the cube method. In *Joint Statistical Meeting of the American Statistical Association*, Seattle August 2006.

Tillé, Y. (2006b). *Sampling Algorithms*. New York: Springer.

Tillé, Y., and Favre, A.-C. (2004). Co-ordination, combination and extension of optimal balanced samples. *Biometrika*, 91, 913-927.

Tillé, Y., and Favre, A.-C. (2005). Optimal allocation in balanced sampling. *Statistics and Probability Letters*, 74, 31-37.

Tillé, Y., and Matei, A. (2007). *The R Package Sampling*. The Comprehensive R Archive Network, http://cran. r-project. org/, Manual of the Contributed Packages.

Tirari, M. (2006). Le plan de sondage équilibré et l'estimation du total d'une population finie. In *Méthodes d'enquêtes et sondages* : *pratiques européenne et nord-américaine*, (Eds., P. Lavallée and L.-P. Rivest), Paris, Dunod, 411-416.

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

Wilms, L. (2000). Présentation de l'échantillon-maître en 1999 et application au tirage des unités primaires par la macro cube. In *Séries Insee Méthodes* : *Actes des Journées de Méthodologie Statistique*, Paris. Insee.

Yates, F. (1946). A review of recent statistical developments in sampling and sampling surveys. *Journal of the Royal Statistical Society*, A109, 12-43.

Yates, F. (1960). *Sampling Methods for Censuses and Surveys*. Charles Griffin, London, England, third edition.

Survey Methodology, December 2011
Vol. 37, No. 2, pp. 227-231
Statistics Canada, Catalogue No. 12-001-X

Innovations in survey sampling design: Discussion of three contributions presented at the U.S. Census Bureau

Jean Opsomer [1]

1. Introduction

The U.S. Census Bureau is one of the largest survey data collection organizations in the world, in addition to its role in the collection of the U.S. Decennial Census data. The two major statistical tools used by the Census Bureau in designing its surveys are stratification and multi-stage sampling. These tools have been successfully implemented starting in the 1940s and have continually been adapted and refined since then.

While this general sampling approach has been very successful, there are increasing concerns about rising survey costs, decreasing response rates and new frame coverage issues (especially related to telephones). At the same time, advances in data collection methods, new data sources and computational tool offer opportunities for considering survey design approaches that would have been unfeasible before. In conjunction with the 2010 Redesign Program currently on-going at the Census Bureau, input was therefore sought from leading academic researchers in innovative sampling methods, as a way to initiate the exploration of possible new approaches to design surveys conducted by the Census Bureau. As a result, Profs. Steve Thompson (Simon Fraser University), Sharon Lohr (Arizona State University) and Yves Tillé (Université de Neufchâtel) were invited to give overview lectures on some of the designs they developed. I was invited to contribute a discussion to each of these lectures.

In the three sections that follow, I will summarize my comments to each of these lectures. My goals in those comments were to highlight the most important aspects of the sampling methods that were presented, to discuss some of the main opportunities for using these designs in the household sampling context, and to identify possible challenges in implementation.

2. Adaptive network and spatial sampling

Prof. Thompson's lecture covered a broad class of designs that includes adaptive cluster sampling, network sampling and adaptive web sampling. Unless I am referring to a specific design within this class, I will refer to these designs as "adaptive sampling" in what follows. A major advantage of adaptive sampling is that it incorporates some of the features of "convenience" sampling approaches such as snowball sampling, including decreased reliance on a sampling frame and the ability to target sampling to portions of the population of particular interest. But unlike convenience sampling, adaptive sampling remains firmly design-based, in the sense of allowing randomization-based finite population estimation and inference.

In adaptive sampling procedures, an initial sample s_0 is drawn according to a probability sampling design $p_0(s_0)$. Based on the characteristics of the elements in s_0 (*e.g.*, presence/absence of features of interest or an enumeration of "links" to other elements in the population), a follow-up sample s_1 is selected from the remaining population, using a conditional sampling design $p_1(s_1 \mid s_0)$. This process is repeated with successive incremental samples $s_2, s_3, \ldots$ until a target criterion such as overall sample size or number of sampling "waves" is reached, and the final sample is the union of each of the successive samples. The specifics on how the waves are drawn varies by adaptive design. Section 2.2 of Thompson's article in this issue and Thompson (2006) provide additional details for adaptive web sampling, a very flexible type of adaptive sampling that includes many of the other designs as special cases.

Because the designs for each of the sampling waves are probability designs, it is possible to obtain valid design-based estimators. A simple estimator for the finite population mean $\mu_N = N^{-1}\sum_U y_i$ is constructed as follows. Based on the initial design p_0 with associated inclusion probabilities π_{0i}, an unbiased estimator for the population mean is given by $\hat{\mu}_0 = N^{-1}\sum_{s_0} y_i / \pi_{0i}$. For each of the subsequent waves $k = 1, \ldots, K$, an unbiased estimator of μ_N is given by $z_k = \sum_{s_{k-1}} y_i + \sum_{s_k} y_i / q_{ki}$, where q_{ki} are conditional inclusion probabilities for wave k (see Thompson (2006) for details on construction of the q_{ki}, and Section 2.4 of Thompson's article in this issue for specific examples). Letting $\hat{\mu}_r = K^{-1}\sum_{k=1}^{K} z_k$, an unbiased estimator for μ_N is obtained as $\hat{\mu} = w\hat{\mu}_0 + (1 - w)\,\hat{\mu}_r$, which is a linear combination of the initial estimator and the mean of the subsequent estimators.

The estimator $\hat{\mu}$ is design unbiased but it depends on the order of the waves in which the sample was obtained. A more precise estimator can be obtained by averaging over all the different orders in which the same sample could have

[1] Jean Opsomer, Department of Statistics, Colorado State University, Fort Collins, CO 80523-1877. E-mail: jopsomer@stat.colostate.edu.

been obtained. For small sample sizes, an explicit expression is available for this more efficient estimator, but in general it needs to be approximated by repeated sampling from an appropriately defined Markov chain, and taking the mean of the samples. The exact methods for setting up the chain and drawing the samples are described in Thompson (2006), which also discusses variance estimation for the resulting estimator.

One of the primary advantages of adaptive sampling designs is that they allow the survey organization to focus the sample in portions of interest in the population. This is particularly useful in situations where some of the elements of interest are relatively rare and where they cannot be identified a priori in a sampling frame. Examples of such situations are surveys of hunting and fishing behavior, recent immigrants, home-schoolers, or owners of family-owned businesses. In each of these cases, the elements are quite "diffuse" in the population and no comprehensive frame is generally available. However, it is likely that individuals who are part of this population will be able to provide information on other individuals, so that links can be identified and sampled across different adaptive sampling waves. Note that adaptive sampling can also be used when these types of rare elements are part of a subpopulation of interest within a survey of a larger and non-rare population. For instance, a survey of school children might want to include a stratum of home-schooled children.

Finding relatively rare (sub)populations is a common challenge in surveys, and a number of methods are regularly deployed to deal with this issue. Perhaps the most common sampling design in the context of household surveys is stratified multi-stage sampling. To the extent that relevant PSU-level auxiliary information is available, the survey organization can oversample PSU expected to contain a larger fraction of the groups of interest. An example of such a situation is a survey of African-American males at risk of Parkinson's disease, in which Census tracts with higher African-American population fraction could be oversampled. Another sampling design that can be useful in this context is multi-phase sampling. In this case, the first phase of sampling is used either as a screening sample or as a way to collect relevant auxiliary information, while subsequent phases focus on obtaining the survey data of interest. The Agricultural Resource Management Survey (ARMS) conducted by the USDA follows this design. A sample of all farms is selected in phase 1, in which farm characteristics for the survey year are collected. In later phases, targeted sampled based on the commodities of interest (*e.g.*, dairy, wheat, *etc*) are selected. A third sampling approach that is sometimes useful for obtaining samples of rare (sub)populations is multi-frame sampling. The principle underlying multi-frame sampling is to combine several frames with different coverage characteristics, for instance a "good" frame containing a large proportion of elements of interest but potentially incomplete and a "bad" frame that is comprehensive but contains a low proportion of elements of interest. For instance, a survey of companies in a particular industry might be able to use an industry group membership list as the "good" frame and a general company list as the "bad" frame. For a more in-depth look at multi-frame sampling, see Section 3 below.

Compared to these three designs, adaptive sampling is more flexible and allows finer control over the number and characteristics of elements that are included in the sample, which will often result in improved efficiency and/or lower cost. A drawback of adaptive sampling is that information needs to be collected on the linkages between elements, which can increase respondent burden and data collection cost, and potentially raises confidentiality issues.

Because adaptive sampling frequently relies on "links" between elements in order to define the conditional selection probabilities in the sampling waves, it is also particularly well-suited for surveys that are interested in studying connections between elements in a population. Examples of such situations might be surveys involving transactions or relationships between businesses, surveys of barter/trading behavior of households, and surveys of family network relationships or characteristics.

For a survey organization contemplating adoption of adaptive sampling, a number of issues related to estimation and data dissemination need to be considered. In many cases, the survey data are released in the form of a weighted dataset, and variance estimates are provided in the form of a simplified design description (*e.g.*, strata and PSUs), replicate weights or generalized variance functions. It is also very common for the weights to be calibrated and/or adjusted for non-response. Estimators for adaptive designs are indeed expressible as weighted sample sums, so that a weighted dataset could readily be created even for the Markov chain version of the estimators mentioned above. The choice of how to best provide variance estimates with the dataset is something that still needs to be investigated and might depend on the specifics of the survey. Similarly, how to incorporate calibration and nonresponse adjustments in adaptive sampling estimation is an area where additional research is needed.

3. Sampling with multiple overlapping frames

Prof. Lohr gave a comprehensive overview of general sampling designs and estimation methods when sampling uses multiple frames. Traditional approaches for conducting surveys are increasingly called into question today, because of increasing costs, decreasing response levels for traditional modes, and increasing concerns for undercoverage of existing sampling frames (*e.g.*, landline telephone numbers reached by RDD). By drawing samples from several frames instead of from a single frame, it is possible to reduce survey costs, improve the coverage of the overall sample, and potentially even increase response rates depending on the specific survey being conducted (for instance, because of improved respondent identifier information in one of the frames).

Multiple frame sampling is a pure randomization-based approach to draw samples, and sampling within the individual frames follows the same methodology as "classical" single-frame sampling. Fully design-based estimation methods for multiple-frame sampling are available, several of which can readily be deployed in the large-scale survey context in which a weighted dataset is the primary output (see below). The key feature of all estimation methods is the estimation of the frame overlap, which is typically unknown but needs to be accounted for. This is done by, for each frame, constructing design-based estimators for the subpopulation(s) of elements that also fall in the other frame(s). The estimators for the characteristics of the frame intersection(s) then need to be combined across frames. Existing methods differ in how they combine these estimators, with the simplest methods using sample-size weighted averages and more complex estimators weighting by estimates of the precision of the individual estimators.

Sampling from multiple frames is particularly applicable in cases where no single frame is available that covers the whole population. Typical examples of such situations are RDD sampling, where an increasing fraction of the population is not reachable through a landline telephone number, surveys of professionals or businesses with partial listings available from vendors or professional organizations. Other situations in which multiple frame sampling might be applicable are surveys of rare subpopulations that exist within a larger population. An overall frame for the population exists, but screening respondents for whether they belong the the subpopulation is time-consuming and expensive. An alternate frame containing a much higher proportion of elements from the subpopulation of interest is sometimes available, but if the coverage of that frame is incomplete, the survey organization might not be willing to rely on it for fear of not obtaining a valid sample. Combining the alternate but incomplete subpopulation frame with the complete but inefficient population frame might be both cost-effective and statistically defensible. Examples of surveys of such subpopulations are surveys of hunting and fishing, where a license frame often exists but it might be incomplete or out of date. This multiple frame approach might also be useful for a survey of the general population, as a way to increase the sample size within certain subpopulations of particular interest. For instance, in a general survey of farms, it might be of interest to produce estimates for organic farms, which only represent a small fraction of farms but with many of those listed in organic business directories. Section 1 of Lohr's article in this issue gives several additional examples of the wide applicability of multiple frame surveys.

As noted above, estimation methods involve the construction of estimators for the frame intersection subpopulation, which requires selection of a weighting method for the estimators obtained from the different frames. Weighting methods that rely on estimating the precision of these estimators might be preferred from an efficiency perspective. However, they are somewhat problematic to implement in practice, because the resulting weights can vary for different variables in the survey. More practical approaches will forego some efficiency in order to be able to have single weights for all survey variables, a key feature emphasized repeatedly in Lohr's article in this issue. The *pseudo-maximum likelihood* (PML) method of Skinner and Rao (1996) produces a single set of weights and is recommended by Lohr as the method of choice for single surveys, while a simpler fixed-weight approach is preferable for longitudinal surveys.

While the basic methodology for constructing design-based estimators for multiple frame sampling is in place today, there is still a need for further research in approaches for applying calibration and nonresponse adjustment in this context. Because it is possible to apply those adjustments at the individual frame level, the population level, or both levels (depending on the available auxiliary information), an investigation of the properties of the estimators under these different scenarios would be very useful, and should be used to develop guidelines for survey practitioners. Section 3 of Lohr's article in this issue discusses some initial results in this area.

Variance estimation methods for multiple-frame estimators have been developed and are reviewed in Section 4.2 of Lohh's article, and include both linearization and replication approaches. An important practical issue in the use of the linearization approach is that it requires access to the frame identification for all the elements in the sample, because it involves separate estimation of the variance in each frame. This might be undesirable for the survey organization producing the data, for reasons of data confidentiality. In the case of replication methods such as jackknife and bootstrap, it is possible for the survey organization to create sets of replicate weights that do not require disclosure of the frame identity of individual sample elements to the data users. Lohr (2007) recommends the *combined bootstrap* approach for inference for multiple frame sampling.

As an alternative, the *grouped jackknife* of Kott (2001) could also be considered.

Implementing multiple frame sampling surveys can be more challenging than single-frame surveys. There needs to be awareness for the increased potential for nonsampling errors, as discussed in Section 5 of Lohr's article, especially if the data collection modes or protocols vary across frames. For instance, sampled elements in one frame get an advance letter, while those in another frame receive a "cold call" because of lack of address information. It is also possible that the nonresponse characteristics differ across frame, so that separate adjustments are required. Finally, in many cases the elements present in the different frames might have different characteristics (*e.g.*, organic farms belonging to a national organic business association *vs.* those that do not). In all those cases, attention to frame-specific effects and careful weight construction are required in order to obtain valid survey estimators. On the other hand, the presence of multiple frames provides opportunities for measuring nonsampling errors, because they entail multiple samples from the same population. For instance, it might be useful to perform "cold calls" for a portion of the selected elements in the frame with addresses to evaluate mode effects.

4. Balanced sampling with the cube method

The presentation by Prof. Tillé covered the fundamentals of balanced sampling and described the *cube method*, which he developed as a practical algorithm implementing the drawing of balanced samples. The goals of balanced sampling designs are to maintain the representation of the population structure in the sample (hence the term "balance"), and to improve the efficiency of survey estimators. Today, most survey statisticians apply stratification as the primary tool to achieve these two goals. Stratification achieves balance by forcing the sample composition to match the stratum allocation, and improves the efficiency of estimators by removing the component of variance due to between-stratum differences. Systematic sampling is also used to achieve these goals, most commonly in natural resource surveys. In this case, the sample composition matches the population composition exactly along the sorting variable, and approximately for any variable correlated with the sorting variable. Efficiency is gained because sample moments of the variables of interest (approximately) match population moments. While both approaches are widely used and work well, they are relatively inflexible. Stratification often involves dividing the population into "cells" defined by the intersection of stratification variables, which might lead to a proliferation of many small cells with corresponding small sample sizes. Systematic sampling is a highly constrained form of sampling with limited amount of flexibility in sample construction, and with the additional issue of the lack of a design-based variance estimator.

Balanced sampling can be viewed as a generalization of stratification. Under this interpretation, stratified samples are drawn with given probabilities of inclusion for all the population elements, but subject to constraints on the sample size in each stratum. In balanced sampling, the stratification constraints are replaced by constraints of the form $\sum_s \boldsymbol{x}_i / \pi_i = \sum_U \boldsymbol{x}_i$, where $\boldsymbol{x}_i$ is a vector of *balancing variables*. When the $\boldsymbol{x}_i$ are stratum indicators, balanced sampling is the same as stratification, but any categorical or continuous variables (or combination thereof) can be used, which provides a high degree of flexibility in sample construction.

As noted above, the cube method is an algorithm that draws balanced samples given a set of inclusion probabilities and constraints. If exactly balanced samples exist in the population, the algorithm will try to select one of them. If no sample can be found that has the postulated inclusion probabilities and satisfies the balancing constraints exactly, it will attempt to come as close as possible to satisfying the constraints. The cube method requires that the balancing variables $\boldsymbol{x}_i$ be known for all elements in the population. Depending on the survey context, this requirement might represent a key limitation on the applicability of balanced sampling.

Despite the fact that balancing on population-level auxiliary variables is done at the design stage, it seems likely that in practice, calibration and other weight adjustments such as for nonresponse will still often be required. In fact, Tillé recommends the combination of balancing and calibration as the most efficient strategy (see Section 7.4 of Tillé's article in this issue). The theoretical properties of estimators that are both balanced and calibrated still needs to be fully worked out, however.

While balanced sampling maintains the inclusion probabilities of the elements in the population, it is clear that the presence of the balancing constraints affects their *joint* inclusion probabilities and hence the variance of the estimators. This topic is addressed in Section 6 of Tillé's article. Deville and Tillé (2005) showed that, under certain conditions, the variance of balanced sampling estimators can be approximated by a linearization-type variance, which depends on the residuals of a linear regression of the survey variables on the balancing variables. While this is an important and useful result, it does not lead to a variance estimation approach that is applicable to all survey applications. One issue is that variance estimation based on this method requires access to the balancing variables for all the survey respondents, and these might not be made publicly available as part of the survey dataset. In this context, a replication-based method might be particularly attractive, because it would not require releasing these variables. However, no such method is currently available.

Balanced sampling has close connections with *rejective sampling*, which aims to achieve the same goals. In rejective sampling, a sample is drawn with prespecified inclusion probabilities and the sample is accepted or rejected based on whether it is within a given tolerance level of a balancing constraint. If the sample is rejected, the procedure is repeated until a sample is found that falls within the tolerance level. While rejective sampling has a long history, Fuller (2009) described some asymptotic theory that showed that asymptotically, his version of rejective sampling was approximately equivalent to balanced sampling.

5. Closing remarks

The methods covered in the three lectures are remarkably complementary. Adaptive designs make it possible to obtain randomization-based, statistically valid samples for populations that have traditionally been difficult to sample efficiently. Very little frame information is required to draw such a sample, but a significant amount of effort has to be expended during the data collection in order to identify and follow the “links” among the elements, and draw the successive samples. In contrast, balanced sampling is useful when very detailed frame information is available, and in that situation, it allows for highly customized and efficient sample designs. Once a balanced sample is drawn, the data collection can proceed in the same manner as for traditional surveys. Multiple frame sampling covers an intermediate case, in the sense that no single good frame exists but several partial frames are used to “offset” each other’s weaknesses. Separate samples are drawn from each frame, and data collection proceeds as usual, except for that fact that it is necessary to determine which frame(s) each sampled respondent belong to.

Combined with the existing approaches already in use, these three new sampling methods have the potential to greatly increase the flexibility with which samples can be customized for specific applications, to reduce survey costs and to increase the precision of survey estimators.

References

Deville, J.-C., and Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 2, 569-591.

Fuller, W.A. (2009). Some design properties of a rejective sampling procedure. *Biometrika*, 96(4), 933-944.

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

Lohr, S. (2007). Recent developments in multiple frame surveys. In *ASA Proceedings of the Joint Statistical Meetings*, American Statistical Association, 3257-3264.

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.

Thompson, S.K. (2006). Adaptive web sampling. *Biometrics*, 62, 1224-1234.

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following people who have provided help or served as referees for one or more papers during 2011.

M. Aitkin, *University of Melbourne*
J.-F. Beaumont, *Statistics Canada*
W. Bell, *U.S. Census Bureau*
E. Berg, *Iowa State University*
Y. Berger, *University of Southampton*
P. Biemer, *Research Triangle Institute*
C. Bocci, *Statistics Canada*
J. van den Brakel, *Statistics Netherlands*
J.M. Brick, *Westat Inc*
J. Bushery, *U.S. Census Bureau*
P. Cantwell, *U.S. Bureau of the Census*
R. Chambers, *Centre for Statistical and Survey Methodology*
R. Creecy, U.S. Census Bureau
M. Di Zio, *Italian National Statistical Institute*
P. Dick, *Statistics Canada*
Y. Dong, *Temple University*
A.H. Dorfman, *U.S. Bureau of Labour Statistics*
G. Dubreuil, *Statistics Canada*
S. Eckman, *Institute for Employment Research*
A. Eideh, *Al-Quds University*
M. Elliott, *University of Michigan*
J.L. Eltinge, *U.S. Bureau of Labor Statistics*
V. Estevao, *Statistics Canada*
G. Forsman, *Swedish Transport Administration*
W.A. Fuller, *Iowa State University*
J. Gambino, *Statistics Canada*
S. Godbout, *Statistics Canada*
H.A. Gutiérrez Rojas, *Faculty of Statistics – Universidad Santo Tomás*
D. Haziza, *Université de Montréal*
Y. He, *Harvard Medical School*
S. Heeringa, *University of Michigan*
M. Hidiroglou, *Statistics Canada*
B. Hulliger, *University of Applied Sciences Northwestern Switzerland*
P.M. Joyce, *U.S. Bureau of Census*
D. Judkins, *Westat Inc*
D. Kasprzyk, *NORC at the University of Chicago*
J.-K. Kim, *Iowa State University*
P. Kott, *Research Triangle Institute*
T. Krenzke, *Westat Inc.*
P. Lahiri, *JPSM, University of Maryland*
P. Lavallée, *Statistics Canada*
H. Lee, *Westat Inc.*
J. Legg, *Amgen Inc., U.S.A.*
J. Li, *Westat, Inc.*
S. Lohr, *Arizona State University*
E. López Escobar, *University of Southampton*
P. Lynn, *Univerity of Essex*
D.J. Malec, *National Center for Health Statistics*
H. Mantel, *Statistics Canada*
K. Miller, *U.S. National Center for Health Statistics*
J.M. Montaquila, *Westat Inc.*
F. Moura, *Universidad Federal de Rio de Janeiro*
T. Mulcahy, *National Opinion Research Center*
J.F. Muñoz Rosas, *University of Granada*
B. Nandram, *WPI*
G. Nathan, *Hebrew University*
D. Nelson, *Center for Chronic Disease Outcomes Research*
S. Oman, *Hebrew University*
J. Opsomer, *Colorado State University*
J.L. Parsons, *USDA, National Agricultural Statistics Service*
Z. Patak, *Statistics Canada*
D. Pfeffermann, *Hebrew University*
C. Poirier, *Statistics Canada*
N.G.N. Prasad, *University of Alberta*
J. Preston, *ABS*
S. Rabe-Hesketh, *UC Berkeley*
J.N.K. Rao, *Carleton University*
J. Reiter, *Duke University*
L.-P. Rivest, *Université Laval*
S. Rubin-Bleuer, *Statistics Canada*
N. Schenker, *National Center for Health Statistics*
F.J. Scheuren, *National Opinion Research Center*
B. Schouten, *Statistics Netherlands*
R. Sigman, *Westat, Inc.*
P. do N. Silva, *Escola Nacional de Ciências Estatísticas*
M. Sinclair, *NORC at the University of Chicago*
A. Singh, *National Opinion Research Center*
B. Skalland, *National Opinion Research Center*
E. Sluid, *University of Maryland*
P. Smith, *Office for National Statistics*
E. Stasny, *Ohio State University*
D. Steel, *University of Wollongong*
L. Stokes, *Southern Methodist University*
M. Sverchkov, *Bureau of Labor Statistics*
M. Thompson, *University of Waterloo*
C. Tucker, Consultant
N. Tzavidis, *University of Southampton*
B. van der Klaauw, *University of Amsterdam*
J. van der Laan, *Statistics Netherlands*
V.J. Verma, *Università degli Studi di Siena*
J. Wang, *Hewlett-Packard Labs*
Z. Wang, *Sir Wilfrid Laurier University*
K.M. Wolter, *National Opinion Research Center*
J. Wood, *Office for National Statistics*
C. Wu, *University of Waterloo*
Y. You, *Statistics Canada*
C. Yu, *Iowa State University*
W. Yung, *Statistics Canada*
A. Zaslavsky, *Harvard Medical School*
L.-C. Zhang, *Statistics Norway*

Acknowledgements are also due to those who assisted during the production of the 2011 issues: Céline Ethier of Statistical Research and Innovation Division, Christine Cousineau and Teresa Jewell of Household Survey Methods Division, Nick Budko and Sophie Chartier of Business Survey Methods Division, Dominique Lavoie of Social Survey Methods Division, Matthew Belyea, Louise Demers, Anne-Marie Fleury, Roberto Guido, Lydia Kokline, Liliane Lanoie, Darquise Pellerin, Joseph Prince and Fadi Salibi of Dissemination Division.

ANNOUNCEMENTS

Nominations Sought for the 2013 Waksberg Award

The journal *Survey Methodology* has established an annual invited paper series in honour of Joseph Waksberg to recognize his contributions to survey methodology. Each year a prominent survey statistician is chosen to write a paper that reviews the development and current state of an important topic in the field of survey methodology. The paper reflects the mixture of theory and practice that characterized Joseph Waksberg's work.

The recipient of the Waksberg Award will receive an honorarium from Westat. The paper will be published in a future issue of *Survey Methodology*.

The author of the 2012 Waksberg paper will be selected by a four-person committee appointed by *Survey Methodology* and the American Statistical Association. Nomination of individuals to be considered as authors or suggestions for topics should be sent before February 28, 2012 to the chair of the committee, Mary Thompson (methomps@uwaterloo.ca).

Previous Waksberg Award honorees and their invited papers are:

2001 Gad **Nathan**, "Telesurvey methodologies for household surveys – A review and some thoughts for the future?". *Survey Methodology*, vol. 27, 1, 7-31.

2002 Wayne A. **Fuller**, "Regression estimation for survey samples". *Survey Methodology*, vol. 28, 1, 5-23.

2003 David **Holt**, "Methodological issues in the development and use of statistical indicators for international comparisons". *Survey Methodology*, vol. 29, 1, 5-17.

2004 Norman M. **Bradburn**, "Understanding the question-answer process". *Survey Methodology*, vol. 30, 1, 5-15.

2005 J.N.K. **Rao**, "Interplay between sample survey theory and practice: An appraisal". *Survey Methodology*, vol. 31, 2, 117-138.

2006 Alastair **Scott**, "Population-based case control studies". *Survey Methodology*, vol. 32, 2, 123-132.

2007 Carl-Erik **Särndal**, "The calibration approach in survey theory and practice". *Survey Methodology*, vol. 33, 2, 99-119.

2008 Mary E. **Thompson**, "International surveys: Motives and methodologies". *Survey Methodology*, vol. 34, 2, 131-141.

2009 Graham **Kalton**, "Methods for oversampling rare subpopulations in social surveys". *Survey Methodology*, vol. 35, 2, 125-141.

2010 Ivan P. **Fellegi**, "The organisation of statistical methodology and methodological research in national statistical offices". *Survey Methodology*, vol. 36, 2, 123-130.

2011 Danny **Pfeffermann**, "Modelling of complex survey data: Why model? Why is it a problem? How can we approach it?". *Survey Methodology*, vol. 37, 2, 115-136.

2012 Lars **Lyberg**, Manuscript topic under consideration.

Members of the Waksberg Paper Selection Committee (2011-2012)

Mary Thompson, *University of Waterloo* (Chair)
J.N.K. Rao, *Carleton University*
Steve Heeringa, *University of Michigan*
Cynthia Clark, *USDA*

Past Chairs:

Graham Kalton (1999 - 2001)
Chris Skinner (2001 - 2002)
David A. Binder (2002 - 2003)
J. Michael Brick (2003 - 2004)
David R. Bellhouse (2004 - 2005)
Gordon Brackstone (2005 - 2006)
Sharon Lohr (2006 - 2007)
Robert Groves (2007 - 2008)
Leyla Mojadjer (2008 - 2009)
Daniel Kasprzyk (2009 - 2010)
Elizabeth A. Martin (2010 - 2011)

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. 2, 2011

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. 3, 2011

All inquires about submissions and subscriptions should be directed to jos@scb.se

CONTENTS TABLE DES MATIÈRES

Volume 39, No. 2, June/juin 2011

The Canadian Journal of Statistics **La revue canadienne de statistique**

CONTENTS TABLE DES MATIÈRES

Volume 39, No. 3, September/septembre 2011

Special Issue: Special Issue in Honour of Jack Kalbfleisch and Jerry Lawless