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

Survey Methodology

Survey Methodology 41-2

Release date: December 17, 2015

Statistics Canada Statistique Canada

Canada

How to obtain more information

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

You can also contact us by

email at STATCAN.infostats-infostats.STATCAN@canada.ca

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

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

Depository Services Program

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

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 "Contact us" > "Standards of service to the public."

Note of appreciation

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

Standard table symbols

The following symbols are used in Statistics Canada publications:

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

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology

Catalogue No. 12-001-XPB

A journal published by Statistics Canada

December 2015

•

Volume 41

•

Number 2

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is indexed in The ISI Web of knowledge (Web of science), The Survey Statistician, Statistical Theory and Methods Abstracts and SRM Database of Social Research Methodology, Erasmus University and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods. It is also covered by SCOPUS in the Elsevier Bibliographic Databases.

EDITORIAL POLICY

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

Submission of Manuscripts

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

Survey Methodology

A Journal Published by Statistics Canada

Volume 41, Number 2, December 2015

Contents

Regular Papers

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.
∞

Vol. 41, No. 2, pp. 267-296
Statistics Canada, Catalogue No. 12-001-X

Dealing with small sample sizes, rotation group bias and discontinuities in a rotating panel design

Jan A. van den Brakel and Sabine Krieg[1]

Abstract

Rotating panels are widely applied by national statistical institutes, for example, to produce official statistics about the labour force. Estimation procedures are generally based on traditional design-based procedures known from classical sampling theory. A major drawback of this class of estimators is that small sample sizes result in large standard errors and that they are not robust for measurement bias. Two examples showing the effects of measurement bias are rotation group bias in rotating panels, and systematic differences in the outcome of a survey due to a major redesign of the underlying process. In this paper we apply a multivariate structural time series model to the Dutch Labour Force Survey to produce model-based figures about the monthly labour force. The model reduces the standard errors of the estimates by taking advantage of sample information collected in previous periods, accounts for rotation group bias and autocorrelation induced by the rotating panel, and models discontinuities due to a survey redesign. Additionally, we discuss the use of correlated auxiliary series in the model to further improve the accuracy of the model estimates. The method is applied by Statistics Netherlands to produce accurate official monthly statistics about the labour force that are consistent over time, despite a redesign of the survey process.

Key Words: Common factor models; Kalman filter; Measurement bias; Small area estimation; Structural time series modelling; Survey sampling.

1 Introduction

Sample surveys of national statistical institutes are generally conducted repeatedly with the purpose of constructing time series that describe the evolution of finite population parameters of interest. Estimation techniques employed by national statistical institutes are largely design based. This implies that statistical inference is predominantly based on the stochastic structure of the sampling design, while statistical models only play a minor role. The general regression (GREG) estimator (Särndal, Swensson and Wretman 1992) is an example of this class of estimators. This estimator expands or weights the observations obtained in the sample with the so-called survey weights, such that the sum over the weighted observations is approximately design unbiased for the unknown population total. The survey weights are initially derived from the sampling design, by taking the weights equal to the inverse of the inclusion probabilities of the sampling units. In a second step these design-weights are calibrated, such that the sum over the weighted auxiliary variables in the sample equates to the known population totals. Under the model-assisted approach, the GREG estimator is derived from a linear regression model that specifies the relationship between the values of a certain target parameter and a set of auxiliary variables.

This class of estimators has nice properties that make them very attractive for use in a production process of compiling timely official statistics. GREG estimators are asymptotically design unbiased and consistent, see Isaki and Fuller (1982), and Robinson and Särndal (1983). This provides a form of robustness in the case of large sample sizes. If the underlying linear model of the GREG estimator explains the variation of the target parameter in the finite population reasonably well, then the use of

[1] Jan A. van den Brakel, Methodology Department, Statistics Netherlands, PO Box 4481, 6401 CZ Heerlen, The Netherlands and Department of Quantitative Economics, Maastricht University School of Business and Economics, PO Box 616, 6200 MD Maastricht, The Netherlands. E-mail: ja.vandenbrakel@cbs.nl; Sabine Krieg, Methodology Department, Statistics Netherlands, PO Box 4481, 6401 CZ Heerlen, The Netherlands.

auxiliary information results in a reduction of the design variance and also decreases the bias due to selective non-response. Model misspecification might result in an increase of the design variance but the property that the GREG estimator is approximately design unbiased remains. From this point of view, the GREG estimator is robust against model misspecification. Additionally, these estimators only require one set of weights to estimate all possible target variables, which is an attractive practical advantage in multipurpose surveys.

Major drawbacks of GREG estimators are the relatively large design variances in the case of small sample sizes, and the fact that they do not handle measurement errors effectively. In such situations, model-based procedures can be used to produce more reliable estimates. These estimators employ sample information observed in other domains or previous time periods through an explicit statistical model, thus increasing the effective sample size in the separate domains and specific periods. In survey methodology, this type of estimation techniques is known as small area estimation, see Rao (2003) for a comprehensive overview. In this paper we describe an estimation approach, based on structural time series modelling, to deal with small sample sizes and problems with non-sampling errors in the Dutch Labour Force Survey (LFS).

Official monthly statistics about the Dutch labour force are based on the Dutch LFS. This survey is based on a rotating panel design. The responding households are interviewed five times at quarterly intervals, which implies that every month five panels are being interviewed. The estimation procedure of the LFS is based on the GREG estimator.

This paper solves three major problems encountered with this survey. The first problem is that the monthly sample size of the LFS is too small to rely on the GREG estimator to produce timely official monthly statistics about the employed and unemployed labour force. Therefore many national statistical institutes publish rolling quarterly figures about the labour force each month. Rolling quarterly figures have the obvious disadvantages that monthly seasonal patterns are smoothed out and that they are less timely since the monthly publications refer to the latest rolling quarter instead of the latest month.

The second problem is that there are substantial systematic differences between the subsequent panels due to mode and panel effects. This is a well-known problem for rotating panel designs, and in the literature this is referred to as rotation group bias (RGB), Bailar (1975). At the moment that the LFS changed from a cross-sectional survey to a rotating panel design in October 1999, the effects of the RGB on the outcomes of the LFS became very visible. This was the direct cause for developing procedures that account for this RGB.

The third problem is the systematic effect on the outcomes of the LFS due to a major redesign of the survey process in 2010. Redesigns generally affect the various non-sampling error sources in a survey process, and therefore result in systematic effects on the outcomes of a survey. In an ideal survey transition process, these so-called discontinuities are quantified in order to keep series consistent and preserve comparability of the outcomes over time. In this redesign the first panel under the old and the new design is conducted in parallel for a period of six months, which provides a direct estimate for the discontinuities in the first panel.

Pfeffermann (1991) proposed a multivariate structural time series model for rotating panels to borrow strength over time and to account for RGB in the level of monthly labour force series. Van den Brakel and Krieg (2009) applied this model to the LFS to estimate the monthly unemployment rate. They extended the model to account for RGB in the level and the seasonal patterns of the monthly unemployment rate series. Van den Brakel and Roels (2010) proposed an intervention analysis approach to estimate discontinuities due to a redesign of cross-sectional surveys, as an alternative for a parallel run.

In this paper, the model proposed by Pfeffermann (1991) is extended with this intervention approach and available auxiliary series. We describe how this model increases the precision of direct estimates by taking advantage of sample information from previous periods, and accounts for the autocorrelation in the sampling errors of the different panels, the RGB, and the discontinuities that arise by the change-over to a new survey process. We focus on how this model enables Statistics Netherlands to publish sufficiently reliable official monthly statistics about the labour force instead of rolling quarterly figures, commonly published by national statistical institutes. We illustrate how the model facilitates a smooth change-over to a new survey design by modelling discontinuities with intervention variables. An important question that will be addressed is how the information from the parallel run in the first panel can be used in the time series model. Finally we illustrate how available auxiliary information about the number of people that are formally registered at the employment office can be incorporated in the time series model to improve the estimates of the discontinuities as well as the precision of the model estimates.

The paper starts in Section 2 with a brief description of the LFS and the problems encountered with the chosen survey design. Section 3 describes the proposed time series model to estimate monthly labour force figures. Section 4 describes the implementation of the time series model before the redesign and compares the results of the time series model with the rolling quarterly figures. The introduction of the new survey design is accompanied by a parallel run of six months, which is described in Section 5. Six different methods are proposed to handle the problems with discontinuities induced by the redesign in Section 6. Results obtained with these methods are compared in Section 7, including a motivation for the method that is finally chosen to produce official statistics. The paper concludes with a discussion in Section 8.

2 Design of the Dutch Labour Force Survey

The objective of the Dutch LFS is to provide reliable information about the Dutch labour force. Each month a stratified two-stage cluster design of addresses is drawn. Strata are formed by geographical regions. Municipalities are considered as primary and addresses as secondary sampling units. All households residing at an address, up to a maximum of three, are included in the sample. Different subpopulations are oversampled to improve the accuracy of the official releases, for example, addresses where people live who are formally registered at the employment office, and subpopulations with low response rates.

Before 2000, the LFS was designed as a cross-sectional survey. Since October 1999, the LFS has been conducted as a rotating panel design. Until the redesign in 2010, data in the first panel were collected by means of computer assisted personal interviewing (CAPI). Respondents were re-interviewed four times at quarterly intervals by means of computer assisted telephone interviewing (CATI). During these re-interviews, a condensed questionnaire was used to establish changes in the labour market position of the respondents. The monthly gross sample size for the first panel averaged about 8,000 addresses commencing the moment that the LFS changed to a rotating panel design and gradually fell to about 6,500 addresses in 2012. The response rate is about 55% in the first panel and in the subsequent panels about 90% with respect to the responding households from the preceding panel.

The estimation procedure of the LFS starts with the GREG estimator. Inclusion probabilities reflect the sampling design and differences in response rates between geographic regions. The weighting scheme is based on a combination of different socio-demographic categorical variables. Key parameters of the LFS are the employed, unemployed and total labour force, which are defined as population totals. Another important parameter is the unemployment rate, which is defined as the ratio of the unemployed labour force to the total labour force.

Figure 2.1 illustrates the RGB for the unemployed labour force. The series of the GREG estimates of the first panel are compared with the average of the GREG estimates of the four subsequent panels. The GREG estimates for the unemployed labour force in the subsequent panels are systematically smaller than in the first panel. The RGB is a consequence of different non-sampling errors like selective non-response, panel attrition, mode-effects, effects due to differences between the CAPI questionnaire and the CATI questionnaire, and panel effects.

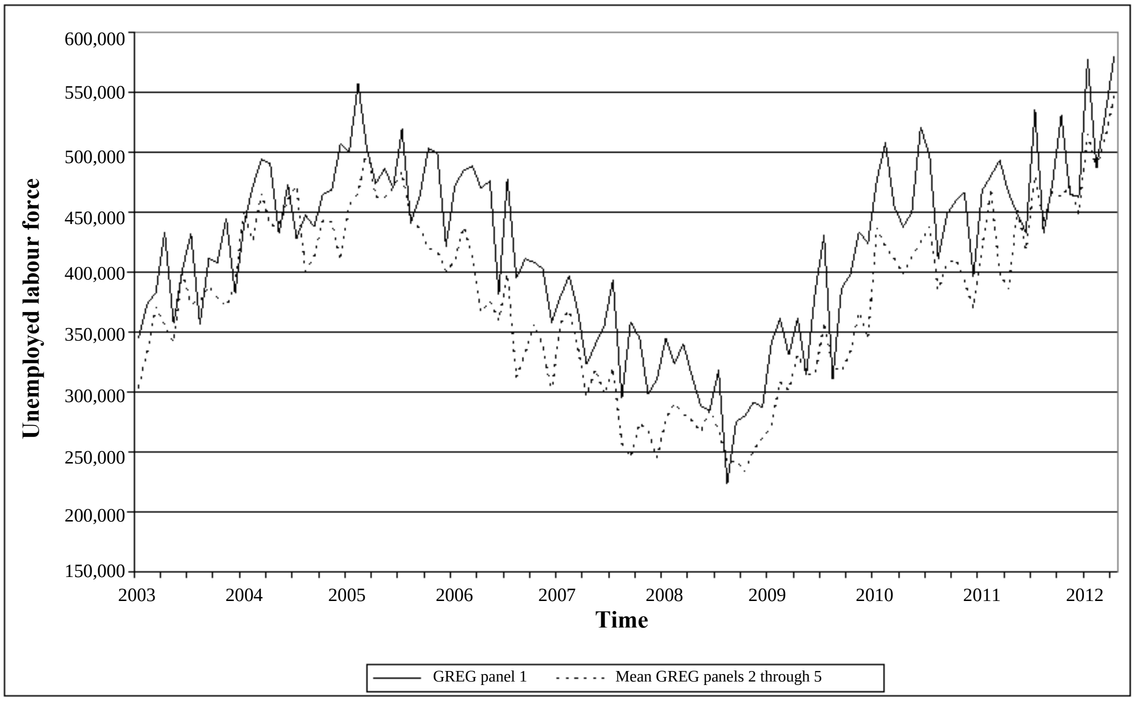


Figure 2.1 RGB unemployed labour force at the national level; comparison GREG estimates based on panel 1 with the mean of the series of the GREG estimates based on panel 2 through 5.

Until June 2010, rolling quarterly figures about the labour force were published each month. A rigid correction was applied to correct for the RGB. For the most important parameters, the ratio between the estimates based on the first panel only and the estimates based on all panels was computed using the data of the 12 preceding quarters. Estimates for the rolling quarterly figures were multiplied by this ratio to correct for RGB. In June 2010, a structural time series model was implemented to estimate model-based monthly figures instead of design-based rolling quarterly figures about the labour force. This model accounts for the RGB, and therefore replaces the ratio correction.

In 2010, a major redesign for the LFS started. The main objective of this redesign was to reduce the administration costs of this survey. This is accomplished by changing the data collection in the first panel from CAPI to a mixed data collection mode using CAPI and CATI. Households with a listed telephone number are interviewed by telephone, the remaining households are interviewed face-to-face. To make CATI data collection in the first panel feasible, the questionnaire for the first panel needed to be abridged since a telephone interview should not take longer than 15 to 20 minutes. Therefore parts of the questionnaire were transferred from the first to the second or the third panel. To avoid confounding real developments with systematic effects induced by the redesign, it is important to quantify these discontinuities and to account for these effects in the time series model.

3 Estimating monthly labour force figures

In this section a multivariate structural time series model is developed for the LFS data that are observed under the rotating panel design. The model deals with small sample sizes by borrowing strength over time to improve the precision of the GREG estimates, and accounts for the RGB as well as the autocorrelation between the subsequent panels of the rotating panel and models the discontinuities due to the redesign of the LFS in 2010.

Let $\hat{Y}_t^j$ denote the GREG estimate for the unknown population parameter, say $\theta_t,$ based on the j^{th} panel observed at time $t, j = 1, \ldots, 5.$ Since responding households are interviewed at quarterly intervals, it follows that the j^{th} panel at time t that was sampled for the first time at time $t - 3j + 3.$ Due to the applied rotation pattern, each month data are collected in five different panels and a vector $\hat{\mathbf{Y}}_t = \left(\hat{Y}_t^1, \hat{Y}_t^2, \hat{Y}_t^3, \hat{Y}_t^4, \hat{Y}_t^5\right)^T$ is observed. A five dimensional time series with GREG estimates for the monthly employed and unemployed labour force is obtained as a result. Pfeffermann (1991) proposed a multivariate structural time series model for this kind of time series to model the population parameter of interest, and to account for the RGB and the autocorrelation in the sampling errors. This approach is extended with an intervention component to model the discontinuities of the survey redesign. This results in the following time series model for the five series of GREG estimates:

$$\hat{\mathbf{Y}}_t = \mathbf{1}_5 \theta_t + \boldsymbol{\lambda}_t + \boldsymbol{\Delta}_t \boldsymbol{\beta} + \mathbf{e}_t, \tag{3.1}$$

with $\mathbf{1}_5$ a five dimensional vector with each element equal to one, $\boldsymbol{\lambda}_t = \left(\lambda_t^1, \lambda_t^2, \lambda_t^3, \lambda_t^4, \lambda_t^5\right)^T$ a vector with time dependent components that account for the RGB, $\boldsymbol{\Delta}_t = \text{Diag}\left(\delta_t^1, \delta_t^2, \delta_t^3, \delta_t^4, \delta_t^5\right)$ a diagonal matrix with dummy variables that change from zero to one at the moment that the survey changes from the old to the new design, $\boldsymbol{\beta} = \left(\beta^1, \beta^2, \beta^3, \beta^4, \beta^5\right)^T$ a five dimensional vector with regression coefficients, and $\mathbf{e}_t = \left(e_t^1, e_t^2, e_t^3, e_t^4, e_t^5\right)^T$ the corresponding survey errors for each panel estimate.

The population parameter θ_t in (3.1) can be decomposed in a trend component, a seasonal component, and an irregular component, i.e.,

$$\theta_t = L_t + S_t + \varepsilon_t. \tag{3.2}$$

Here L_t denotes a stochastic trend component, using the so-called smooth trend model,

$$\begin{aligned} L_t &= L_{t-1} + R_{t-1}, \\ R_t &= R_{t-1} + \eta_t, \\ E(\eta_t) &= 0, \quad \operatorname{Cov}(\eta_t, \eta_{t'}) = \begin{cases} \sigma_\eta^2 & \text{if } t = t' \\ 0 & \text{if } t \neq t'. \end{cases} \end{aligned} \tag{3.3}$$

A likelihood ratio test indicates that in this application the more general local linear trend model, which has a disturbance term for the slope parameter R_t as well as a disturbance term for the level parameter L_t, does not improve the fit to the data. Inclusion of a disturbance term for the level increases the log-likelihood of (3.1) with 0.05 units. This results in a likelihood ratio test statistic of 0.1. Under the null hypothesis that the level disturbance term is equal to zero, this test statistic is a chi-squared distributed random variable with 1 degree of freedom. As a result, this null hypothesis is accepted with a $p -$value of 0.75.

Furthermore, S_t denotes a trigonometric stochastic seasonal component,

$$S_t = \sum_{l=1}^{6} S_{l,t}, \tag{3.4}$$

where

$$\begin{aligned} S_{l,t} &= S_{l,t-1} \cos(h_l) + S_{l,t-1}^* \sin(h_l) + \omega_{l,t} \\ S_{l,t}^* &= S_{l,t-1}^* \cos(h_l) - S_{l,t-1} \sin(h_l) + \omega_{l,t}^*, \quad h_l = \frac{\pi l}{6}, \quad l = 1, \ldots, 6, \\ E(\omega_{l,t}) &= E(\omega_{l,t}^*) = 0, \\ \operatorname{Cov}(\omega_{l,t}, \omega_{l',t'}) &= \operatorname{Cov}(\omega_{l,t}^*, \omega_{l',t'}^*) = \begin{cases} \sigma_\omega^2 & \text{if } l = l' \text{ and } t = t' \\ 0 & \text{if } l \neq l' \text{ or } t \neq t' \end{cases}, \\ \operatorname{Cov}(\omega_{l,t}, \omega_{l,t}^*) &= 0, \quad \forall l, \forall t. \end{aligned} \tag{3.5}$$

Finally, ε_t denotes the irregular component, which contains the unexplained variation of the population parameter and is modelled as a white noise process:

$$E(\varepsilon_t) = 0, \quad \operatorname{Cov}(\varepsilon_t, \varepsilon_{t'}) = \begin{cases} \sigma_\varepsilon^2 & \text{if } t = t' \\ 0 & \text{if } t \neq t'. \end{cases} \tag{3.6}$$

It is not immediately obvious that the white noise component ε_t in (3.2) and the sampling errors $\mathbf{e}_t$ in (3.1) are both identifiable. The sampling errors can be separated from the white noise component because each sample is observed five times and because the variance of the sampling errors, as well as the autocorrelation in the sampling errors induced by the sample overlap of the panel, are calculated directly from the survey data. Details are explained below.

The trend (3.3) describes the gradual change in the population parameter, while the seasonal component (3.4) captures the systematic monthly deviations from the trend within a year. See e.g., Durbin and Koopman (2001) for details. Through component (3.2) values for θ_t are related to the population values from preceding periods. This component shows how sample information observed in preceding periods is used to improve the precision of the estimates for θ_t in a particular time period.

The systematic differences between the subsequent panels, i.e., the RGB, are modelled in (3.1) with $\boldsymbol{\lambda}_t$. The absolute bias in the monthly labour force figures cannot be estimated from the sample data only. Therefore additional restrictions for the elements of $\boldsymbol{\lambda}_t$ are required to identify the model. Here it is assumed that an unbiased estimate for θ_t is obtained with the first panel, i.e., $\hat{Y}_t^1$. This implies that the first component of $\boldsymbol{\lambda}_t$ equals zero. The other elements of $\boldsymbol{\lambda}_t$ measure the time dependent differences with respect to the first panel. Contrary to Pfeffermann (1991), were time independent RGB is assumed, λ_t^j are modelled as random walks for $j = 2, 3, 4,$ and $5.$ As a result it follows that

$$\lambda_t^1 = 0, \quad \lambda_t^j = \lambda_{t-1}^j + \eta_{\lambda,j,t}, j = 2, 3, 4, 5, \tag{3.7}$$

$$E\left(\eta_{\lambda,j,t}\right) = 0, \quad \operatorname{Cov}\left(\eta_{\lambda,j,t}, \eta_{\lambda,j',t'}\right) = \begin{cases} \sigma_\lambda^2 & \text{if} \quad t = t' \quad \text{and} \quad j = j' \\ 0 & \text{if} \quad t \neq t' \quad \text{or} \quad j \neq j'. \end{cases}$$

The discontinuities induced by the redesign in 2010 are modelled with the third term in (3.1). The diagonal matrix $\boldsymbol{\Delta}_t$ contains five intervention variables:

$$\delta_t^j = \begin{cases} 0 & \text{if} \quad t < T_R^j \\ 1 & \text{if} \quad t \geq T_R^j \end{cases}, \text{for } j = 1, 2, \ldots, 5, \tag{3.8}$$

where T_R^j denotes the moment that panel j changes from the old to the new survey design. Under the assumption that (3.2) correctly models the evolution of the population variable, the regression coefficients in $\boldsymbol{\beta}$ can be interpreted as the systematic effects of the redesign on the level of the series observed in the five panels. The intervention approach with state-space models was originally proposed by Harvey and Durbin (1986) to estimate the effect of seat belt legislation on British road casualties. With step intervention (3.8) it is assumed that the redesign only has a systematic effect on the level of the series. Alternative interventions, e.g., for the slope or the seasonal components are also possible, see Durbin and Koopman (2001), Chapter 3. A redesign might not only affect the point estimates, but also the variance of the GREG estimates. This issue is discussed under the time series model for the survey errors.

Finally a time series model for the survey errors $\mathbf{e}_t$ in (3.1) is developed. The direct estimates for the design variances of the survey errors are available from the micro data and are incorporated in the time series model using the survey error model $e_t^j = k_t^j \tilde{e}_t^j$ where $k_t^j = \sqrt{\hat{\operatorname{Var}}\left(\hat{Y}_t^j\right)},$ proposed by Binder and Dick (1990). Here $\hat{\operatorname{Var}}\left(\hat{Y}_t^j\right)$ denotes the estimated variance of the GREG estimator. Choosing the survey errors proportional to the standard error of the GREG estimators allows for non-homogeneous variance in the survey errors, that arise e.g., due to the gradually decreasing sample size over the last decade.

The sample of the first panel has no sample overlap with panels observed in the past. Consequently, the survey errors of the first panel, e_t^1, are not correlated with survey errors in the past. It is, therefore, assumed that $\tilde{e}_t^1$ is white noise with $E\left(\tilde{e}_t^1\right) = 0$ and $\operatorname{Var}\left(\tilde{e}_t^1\right) = \sigma_{e1}^2$. As a result, the variance of the survey error equals $\operatorname{Var}\left(e_t^1\right) = \left(k_t^1\right)^2 \sigma_{e1}^2$, which is approximately equal to the direct estimate of the variance of the GREG estimate for the first panel if the maximum likelihood (ML) estimate for σ_{e1}^2 is close to one.

The survey errors of the second, third, fourth and fifth panel are correlated with survey errors of preceding periods. The autocorrelations between the survey errors of the subsequent panels are estimated from the survey data, using the approach proposed by Pfeffermann, Feder and Signorelli (1998). In this application it appears that the autocorrelation structure for the second, third, fourth and fifth panel can be modelled conveniently with an AR(1) model, van den Brakel and Krieg (2009). Therefore it is assumed that $\tilde{e}_t^j = \rho \tilde{e}_{t-3}^{j-1} + \nu_t^j$, with ρ the first order autocorrelation coefficient, $E\left(\nu_t^j\right) = 0$, and $\operatorname{Var}\left(\nu_t^j\right) = \sigma_{ej}^2$ for $j = 2, 3, 4, 5$. Since $\tilde{e}_t^j$ is an AR(1) process, $\operatorname{Var}\left(e_t^j\right) = \sigma_{ej}^2 \left(k_t^j\right)^2 \big/ \left(1 - \rho^2\right)$. As a result $\operatorname{Var}\left(e_t^j\right)$ is approximately equal to $\operatorname{V\hat{a}r}\left(\hat{Y}_t^j\right)$ provided that the ML estimates for σ_{ej}^2 are close to $\left(1 - \rho^2\right)$.

The survey redesign in 2010 might affect the variance of the GREG estimates. Systematic differences in these variances are automatically taken into account, since they are used as a-priori information in the time series model for the survey error. An alternative possibility would be to allow for different values for σ_{ej}^2 before and after the survey redesign, which can be interpreted as an intervention on the variance hyperparameter of the survey error.

Auxiliary time series can be incorporated in the model to improve the estimates for the discontinuities. Reliable auxiliary series contain valuable information for correctly separating real developments from discontinuities in the intervention model. The auxiliary information will also increase the precision of the model estimates for the monthly unemployment figures. For the unemployed labour force, the number of people formally registered at the employment office is a potential auxiliary variable to be included in the model.

There are different ways to incorporate auxiliary information in the model. One straightforward possibility is to extend the time series model (3.2) for the population parameter of the LFS with a regression component for the auxiliary series, i.e., $\theta_t = L_t + S_t + bX_t + \varepsilon_t$, where X_t denotes the auxiliary series and b the regression coefficient. The major drawback of this approach is that the auxiliary series will partially explain the trend and seasonal effect in θ_t, leaving only a residual trend and seasonal effect for L_t and S_t. This hampers the estimation of a trend for the target variable.

An alternative approach, that allows the direct estimation of a filtered trend for θ_t, is to extend model (3.1) with the auxiliary series and model the correlation between the trends of the series of the LFS and the auxiliary series. This gives rise to the following model:

$$\begin{pmatrix} \mathbf{Y}_t \\ X_t \end{pmatrix} = \begin{pmatrix} \mathbf{1}_5 \theta_t^{\mathrm{LFS}} \\ \theta_t^R \end{pmatrix} + \begin{pmatrix} \boldsymbol{\lambda}_t \\ 0 \end{pmatrix} + \begin{pmatrix} \boldsymbol{\Delta}_t \boldsymbol{\beta} \\ 0 \end{pmatrix} + \begin{pmatrix} \mathbf{e}_t \\ 0 \end{pmatrix}. \tag{3.9}$$

The series of the LFS and the auxiliary series from the register both have their own population parameter that can be modelled with two separate time series models, i.e., $\theta_t^z = L_t^z + S_t^z + \varepsilon_t^z,$ where $z = \text{LFS}$ or $z = R$ (R stands for register), defined similarly to (3.2). Since the auxiliary series is based on a registration, this series does not have a RGB, a discontinuity at the moment that the LFS is redesigned or a survey error component.

The model allows for correlation between the disturbances of the slope of the trend component of the LFS and the auxiliary series. This results in the following definition for the smooth trend model for the LFS and the auxiliary series:

$$\begin{aligned}
L_t^z &= L_{t-1}^z + R_{t-1}^z, \\
R_t^z &= R_{t-1}^z + \eta_t^z, \\
E\left(\eta_t^z\right) &= 0, \\
\operatorname{Cov}\left(\eta_t^z, \eta_{t'}^z\right) &= \begin{cases} \sigma_{\eta^z}^2 & \text{if} \quad t = t' \\ 0 & \text{if} \quad t \neq t' \end{cases}, \quad z = \text{LFS}, R, \\
\operatorname{Cov}\left(\eta_t^{\text{LFS}}, \eta_{t'}^R\right) &= \begin{cases} \vartheta \sigma_{\eta \text{LFS}} \sigma_{\eta R} & \text{if} \quad t = t' \\ 0 & \text{if} \quad t \neq t', \end{cases}
\end{aligned}$$

with ϑ the correlation coefficient between these series. The correlation between both series is determined by the model. If the model detects a strong correlation, then the trends of both series will develop into the same direction more or less simultaneously. Model (3.9) does not allow for correlation between the disturbances of the seasonal component of the LFS series and the auxiliary series. Both series have their own seasonal component S_t^z defined by (3.5). In a similar way both series have their own white noise ε_t^z for the unexplained variation, which are assumed to be uncorrelated and are defined by (3.6).

Models (3.1) and (3.9) explicitly account for discontinuities in the different panels through the intervention component. Estimates for the target variables, obtained with these models, are therefore not affected by the systematic effect of the change-over. As a result, the models correct for the discontinuities induced by the redesign. Model estimates for the target variables can be interpreted as the results observed under the old method, also after the change-over to the new survey design. The discontinuity of the first panel must be added to the model estimates for the target variables to produce figures that can be interpreted as being obtained under the new design.

The general way to proceed is to express the model in the so-called state-space representation and apply the Kalman filter to obtain optimal estimates for the state variables, see e.g., Durbin and Koopman (2001). It is assumed that the disturbances are normally distributed. Under this assumption, the Kalman filter gives optimal estimates for the state vector and the signals. Estimates for state variables for period t based on the information available up to and including period t are referred to as the filtered estimates. The filtered estimates of past state vectors can be updated if new data become available. This procedure is referred to as smoothing and results in smoothed estimates that are based on the completely observed time series. In this application, interest is mainly focussed on the filtered estimates, since they are based on the complete set of information that would be available in the regular production process to produce a model-based estimate for month t.

The analysis is conducted with software developed in OxMetrics in combination with the subroutines of SsfPack 3.0, see Doornik (2009) and Koopman, Shephard and Doornik (2008). All state variables are non-stationary with the exception of the survey errors. The non-stationary variables are initialised with a diffuse prior, i.e., the expectation of the initial states is equal to zero and the initial covariance matrix of the states is diagonal with large diagonal elements. The survey errors are stationary and therefore initialised with a proper prior. The initial values for the survey errors are equal to zero and the covariance matrix is available from the aforementioned model for the survey errors. In Ssfpack 3.0 an exact diffuse log-likelihood function is obtained with the procedure proposed by Koopman (1997).

4 Implementation

In this section we compare the results obtained with the time series model with the GREG estimator for the period before the change-over to the new design, since rolling quarterly data are not calculated during and after the implementation of the new design. Since June 2010 model (3.1) has been applied to produce official monthly figures about the unemployed labour force, the employed labour force and the total labour force at the national level, and for six domains (men and women in three age classes). The model is applied to each variable separately. Estimates are computed as the sum of the trend and the seasonal effects, which is further referred to as the signal. Furthermore, trend estimates are published, replacing previous seasonally corrected figures. The first years of the GREG series are used to obtain stable estimates for the state variables of model (3.1). At the moment of implementation, a series of monthly figures starting in January 2003 is published.

Table 4.1 provides an overview of the ML estimates of the hyperparameters and the autocorrelation in the survey errors. The assumptions underlying the state-space model are evaluated by testing whether the standardized innovations are standard normally and independently distributed, Durbin and Koopman (2001), Section 4.2.4. Bowman-Shenton normality tests, $\mathrm{F}-$tests for heteroscedasticity, $\mathrm{QQ}-$plots, plots of standardized innovations and sample correlograms indicate that these assumptions are not violated under model (3.1).

Table 4.1
ML estimates of hyperparameters for monthly unemployed labour force figures before the survey redesign. Values are expressed as standard deviations

Standard deviation		National level	Men 15-24	Women 15-24	Men 25-44	Women 25-44	Men 45-64	Women 45-64
Slope	$(\hat{\sigma}_\eta)$	2,079	248	179	724	463	412	228
Seasonal	$(\hat{\sigma}_\omega)$	0.02	0.00	0.00	0.00	0.00	0.04	0.22
RGB	$(\hat{\sigma}_\lambda)$	905	941	468	268	669	3	335
White noise	$(\hat{\sigma}_\varepsilon)$	6,884	1,528	3,521	4,359	4,294	3,329	2
Survey error panel 1	$(\hat{\sigma}_{e1})$	1.07	0.98	1.11	1.04	0.89	0.99	1.14
Survey error panel 2	$(\hat{\sigma}_{e2})$	0.99	0.95	1.03	1.03	0.94	1.17	1.02
Survey error panel 3	$(\hat{\sigma}_{e3})$	1.01	1.06	1.12	1.03	0.96	1.04	0.92
Survey error panel 4	$(\hat{\sigma}_{e4})$	1.13	1.07	1.21	0.99	0.96	0.99	0.95
Survey error panel 5	$(\hat{\sigma}_{e5})$	1.06	1.00	1.03	0.99	0.99	1.08	0.87
Autocorrelation	$(\hat{\rho})$	0.21	0.13	0.12	0.39	0.22	0.44	0.38

The hyperparameter estimates for the survey errors for panel 2, 3, 4 and 5 are divided by $\left(1-\hat{\rho}^2\right)$. Therefore hyperparameters for the survey errors are, as expected, around 1.

In Figure 4.1, the filtered estimates for the monthly unemployed labour force at the national level based on model (3.1) are compared with the monthly GREG estimates and with the rolling quarterly GREG figures. Both GREG estimates are corrected for RGB using the ratio correction described in Section 2. The three series are at the same level, since they are calibrated to the level of the first panel. The series of the monthly GREG estimates has more pronounced peaks and dips than the filtered estimates. Under the times series model these fluctuations are partially considered as survey errors and filtered from the GREG estimates. The rolling quarterly figures have a less pronounced seasonal pattern, since monthly patterns are averaged over three subsequent months.

Figure 4.2 compares the filtered trend estimates with the seasonally adjusted estimates of the rolling quarterly data for the unemployed labour force at the national level. The seasonally adjusted rolling quarterly data, computed by X-12-ARIMA (U.S. Census Bureau 2009), were published before the new estimation method was implemented, and are available until May 2010. They are computed as the original estimates minus the seasonal effects. Besides the trend, they also include the sampling errors and other irregularities. Seasonally adjusted rolling quarterly figures and the filtered trend therefore measure slightly differently defined concepts. After the implementation of the time series model, the seasonally adjusted figures are replaced by the filtered trend, so it is interesting to compare the differences between both figures mainly to judge how large the consequences are for the users of these data.

There are some minor differences in the levels of the series in Figures 4.1 and 4.2. They are the result of large sampling errors and differences between the methods used to account for RGB. Firstly, the monthly GREG estimates and the rolling quarterly GREG estimates are more sensitive to large sampling errors. This in contrast with the time series model that filters the survey errors from the GREG estimates.

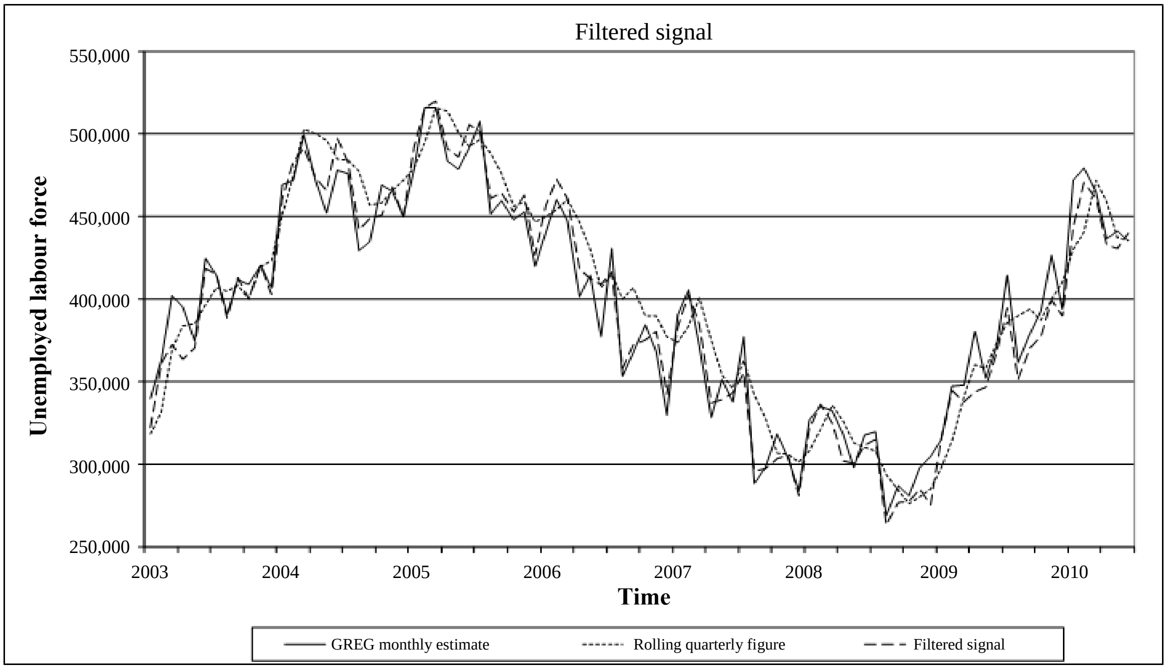


Figure 4.1 **Monthly GREG estimates, rolling quarterly GREG estimates and monthly filtered model estimates, unemployed labour force at the national level.**

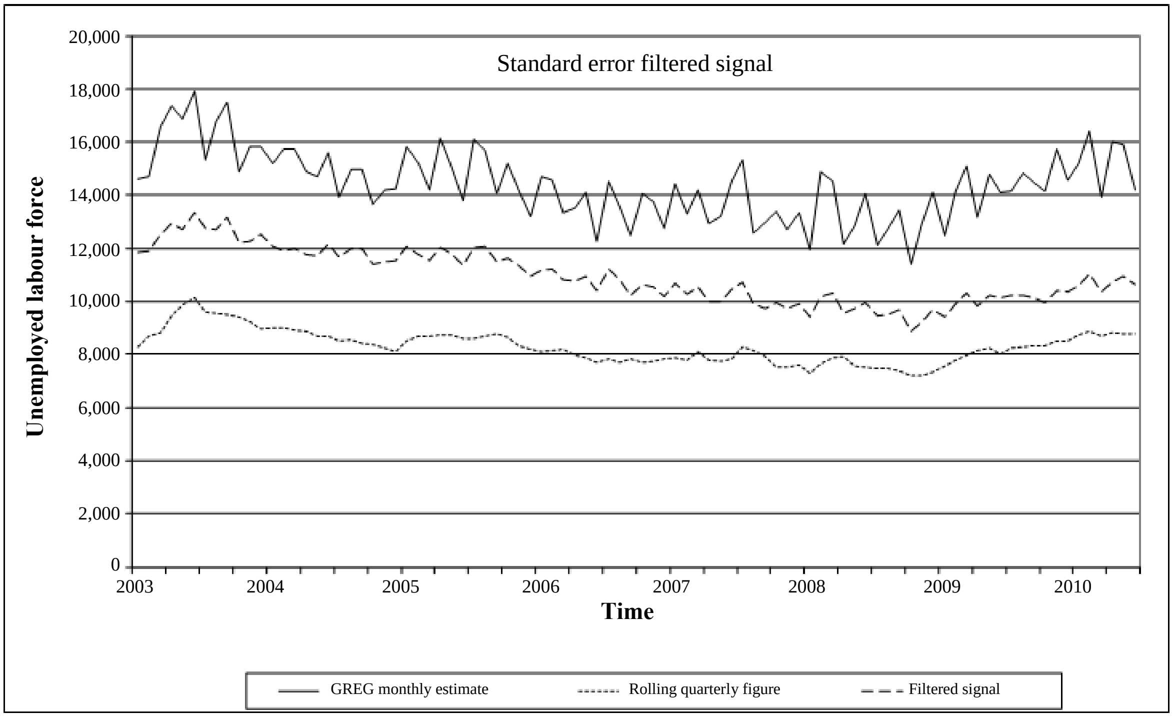


Figure 4.1 (cont.) **Standard errors monthly GREG estimates, rolling quarterly GREG estimates and monthly filtered model estimates, unemployed labour force at the national level.**

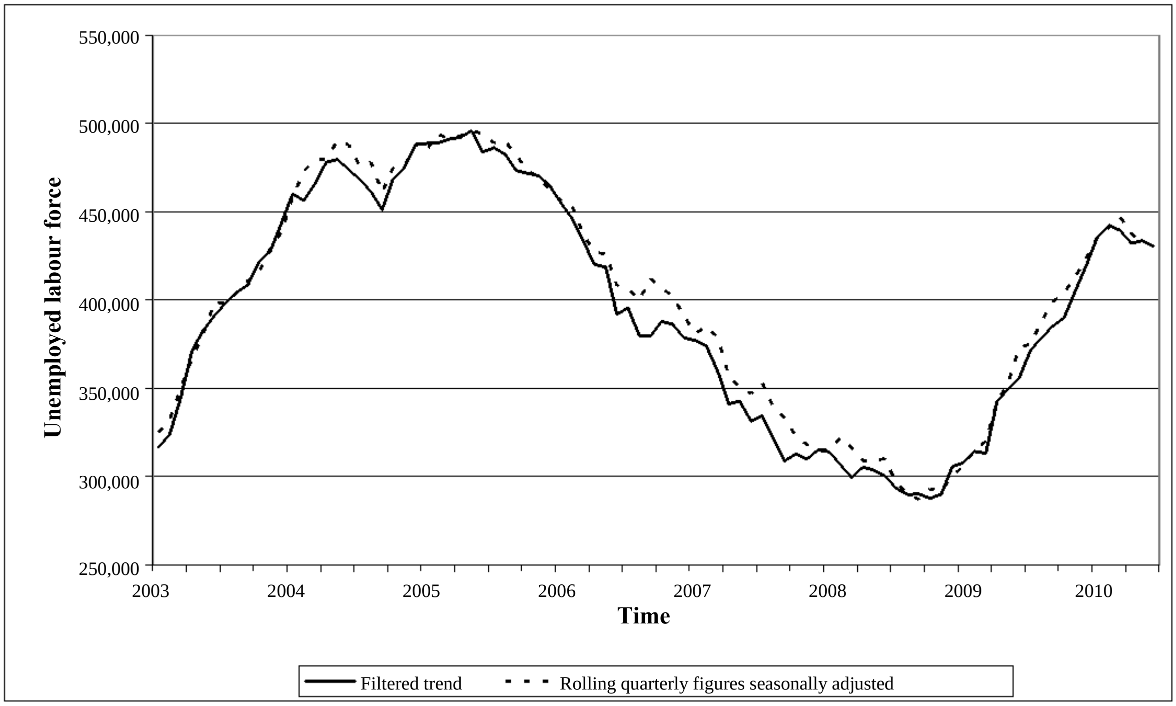


Figure 4.2 **Seasonally adjusted rolling quarterly figures and monthly filtered trend estimates, unemployed labour force at national level.**

Secondly, the RGB correction for the monthly GREG estimates and the rolling quarterly figures are based on a rigid and untested assumption of a constant ratio over a period of three years, see Section 2. In the time series model, the RGB is modelled as differences between the panels and is allowed to change gradually over time, see equation (3.7). Filtered estimates for the RGB in the monthly unemployed labour force at national level are plotted in Figure 4.3. This figure shows that the assumption of a constant ratio over a period of three years is not tenable, since the absolute value of the RGB increases in a period that the unemployed labour force decreases. It is therefore unlikely that the ratio used to correct the rolling quarterly figures is constant over three year periods. The model evaluation does not indicate that the assumptions underlying time series model (3.1) are not met. It can therefore be expected that a more reliable RGB correction is obtained with the time series modelling approach.

Thirdly, the methodology of X-12-ARIMA assumes that there is no autocorrelation in the sampling errors. This assumption is clearly not met in a rotating panel. Pfeffermann et al. (1998) showed that the use of X-12-ARIMA to series with autocorrelated survey errors results in spurious trend estimates. This partially explains the differences between the filtered trend and the seasonally adjusted rolling quarterly data in Figure 4.2.

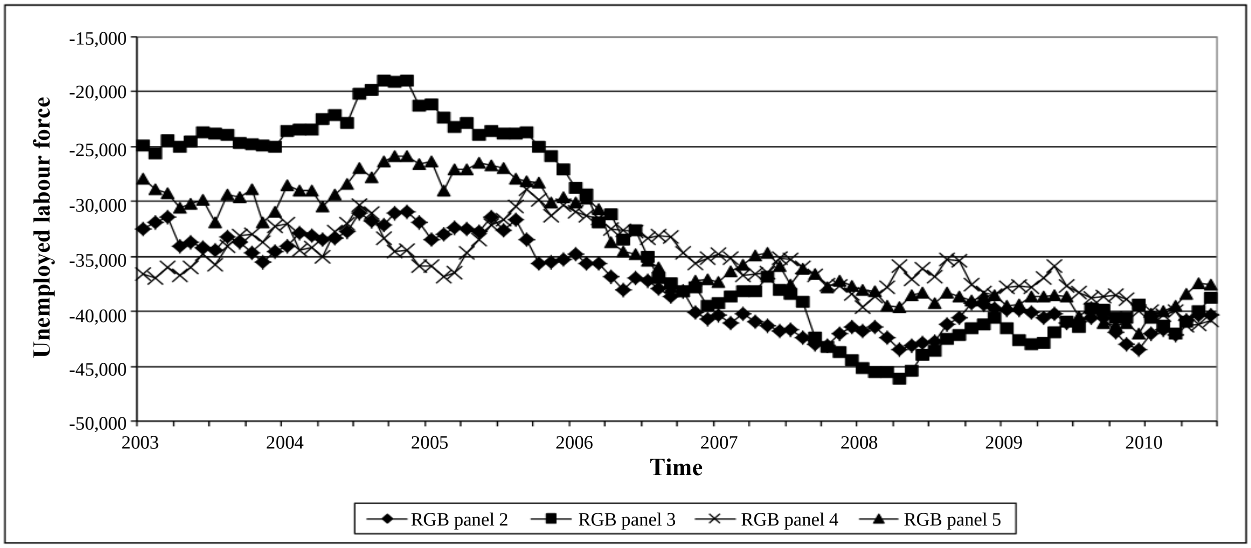


Figure 4.3 Filtered estimates for RGB in the monthly unemployed labour force at national level.

The standard errors of the monthly GREG estimates and the rolling quarterly figures are based on the variance of the Taylor approximation of the GREG estimator, Särndal et al. (1992), Chapter 6. The ratio used to correct for RGB is assumed to be known, although it is based on the samples of three years. The standard errors of the filtered estimates ignore the uncertainty of using ML estimates for the hyperparameters. Table 4.2 compares the means of the standard errors over the last 24 months for the three considered methods for the unemployed labour force, at the national level and for the six domains. Figure 4.1 compares the standard errors at the national level for the three methods for the entire series. In all cases, the precision of the monthly GREG estimates has been substantially improved by the time series model. The rolling quarterly figures have smaller standard errors than the model estimates in almost all cases. For the domains men $15-24$ and women $45-64,$ the precision of the model estimates and of the rolling quarterly figures are similar. Nevertheless, the time series model produces sufficiently reliable monthly estimates to replace the rolling quarterly figures by monthly figures. This circumvents the aforementioned disadvantages of the rolling quarterly figures. Moreover it is not straightforward how rolling quarterly figures can be corrected for RGB in combination with discontinuities induced by the redesign in 2010.

Table 4.2
Mean standard errors unemployed labour force over 24 months (July 2008 – June 2010)

	National level	Men 15-24	Women 15-24	Men 25-44	Women 25-44	Men 45-64	Women 45-64
Rolling quarterly estimate	8,118	3,126	2,831	4,041	3,809	3,452	3,260
Monthly GREG estimate	14,172	5,448	4,885	7,083	6,662	6,046	5,676
Model estimate	10,082	3,247	3,439	5,075	4,749	4,119	3,269
Ratio model and rolling quarterly figure	1.24	1.04	1.21	1.26	1.25	1.19	1.00
Ratio model and monthly GREG estimate	0.71	0.60	0.70	0.72	0.71	0.68	0.58

An artefact of applying model (3.1) to each variable and domain separately is that the sum over the domain estimates is not exactly equal to the estimate at the national level and that the sum of the employed and unemployed labour force is not exactly equal to the total labour force for each domain and at the national level. With the GREG estimator these estimates are consistent by definition, since one set of weights is used to compile all required estimates. The aforementioned restrictions for the model estimates are restored through an appropriate Lagrange function, which distributes the discrepancies over the model estimates proportional to their MSE estimates. Details are given in the Appendix. Finally, unemployment rates are obtained as the ratio of the model estimate for the unemployed labour force to the total labour force for the six domains and the national level.

The model-based domain estimates for the monthly employed and unemployed labour force are included as a weighting term in the GREG estimator for the quarterly and yearly releases. This enforces consistency between monthly, quarterly, and yearly labour force figures and corrects for the RGB in the GREG estimates of the quarterly and yearly labour force figures.

5 Redesign of the Dutch Labour Force Survey

The LFS was redesigned in 2010, as described in Section 2. Discontinuities induced by this redesign were quantified by conducting the first panel under the old and new design in parallel for a period of six months, from January through June 2010. Each month two separate samples with the regular monthly sample size were drawn from the target population according to the sample design of the LFS. One sample was assigned to the old and one to the new LFS design. This made a direct estimate possible for the discontinuities for the main parameters in the first panel.

Mainly due to budget constraints, the subsequent panels were not conducted in parallel under the old and the new design. Possible discontinuities were quantified using the intervention approach described in Section 3. In the time series model, the outcomes of the subsequent panels are benchmarked to the level of the first panel. It is therefore crucial that the first panel is measured as accurately as possible, including possible discontinuities due to a redesign. Therefore it was decided to conduct a sufficiently large parallel run for the first panel, and use the intervention approach for the remaining panels. The estimates for the discontinuities from the parallel run as well as the intervention variables of the time series model are the effect of all factors that changed simultaneously in the redesign of the survey.

In the parallel run, 19,150 responding households under the old design and 16,906 responding households under the new design were obtained. Table 5.1 compares the field work results of the new and old design, both for households with and without a listed phone number. Overall, the response rate is lower for households without a listed phone number. This can be explained by the fact that this part of the population typically consists of hard to reach groups like young people and migrants. Furthermore, the response rate is lower under CATI than under CAPI for households with a listed phone number. Both the percentages of no contact and of frame errors increase substantially when using CATI instead of CAPI. Frame errors under CAPI are mostly non-existing or unoccupied addresses, under CATI they are mostly closed phone lines. Other non-response includes, for example, illness.

Table 5.2 summarizes the estimation results of the parallel run for the unemployed labour force. At the national level, the change-over to the new design resulted in an increase of about 55,000 in the monthly unemployed labour force figures. The differences fluctuated considerably over the six months of the parallel run, probably caused by the large sampling errors of the GREG estimates. A strong increase in the differences was observed in the last two months of the parallel run, particularly at the national level. This can be explained partially by the low response under the new design during these two months.

The decision was made to produce official monthly figures using the data obtained under the old design until June 2010. After completion of the parallel run, all the available data obtained under the new design were used to compile official monthly figures. So since July 2010, the data in the first panel have been based on the new design from January 2010, while the data in the second panel are based on the new design from April 2010, and the data in the third panel are based on the new design from July 2010 and so on.

Table 5.1
Overview fieldwork results of the parallel run first panel

OLD

	CAPI - phone		**CAPI – no phone**		**total**	
Category	households	%	households	%	households	%
Total	20,813	100.0%	14,469	100.0%	35,282	100.0%
Frame errors	769	3.7%	1,039	7.2%	1,808	5.1%
Not approached	618	3.0%	463	3.2%	1,081	3.1%
Language problems	390	1.9%	878	6.1%	1,268	3.6%
Refusal	4,909	23.6%	3,112	21.5%	8,021	22.7%
No contact	889	4.3%	1,455	10.1%	2,344	6.6%
Other non-response	921	4.4%	689	4.8%	1,610	4.6%
Complete response	12,317	59.2%	6,833	47.2%	19,150	54.3%

NEW

	CATI		**CAPI**		**total**	
Category	households	%	households	%	households	%
Total	20,234	100.0%	13,345	100.0%	33,579	100.0%
Frame errors	1,539	7.6%	982	7.4%	2,521	7.5%
Not approached	1	0.0%	428	3.2%	429	1.3%
Language problems	317	1.6%	788	5.9%	1,105	3.3%
Refusal	4,545	22.5%	2,903	21.8%	7,448	22.2%
No contact	2,233	11.0%	1,333	10.0%	3,566	10.6%
Other non-response	963	4.8%	641	4.8%	1,604	4.8%
Complete response	10,636	52.6%	6,270	47.0%	16,906	50.3%

To analyse differences in response distributions between the old and the new design, results must be compared column-wise.

Table 5.2
Comparison of GREG estimates new and old design for monthly unemployed labour force figures, first panel (×1,000), standard errors in brackets, significant difference at a 5% significance level indicated with *

	National level	Men 15-24	Women 15-24	Men 25-44	Women 25-44	Men 45-64	Women 45-64
Monthly unemployed labour force new design – mean over January-June							
	475	67	56	103	101	80	68
Difference new and old design monthly unemployed labour force							
Mean January – June	55*(17)	19*(6)	7 (6)	-1 (9)	20*(8)	6 (8)	4 (7)
Difference per month							
January	56 (39)	13 (14)	1 (14)	-15 (21)	-16 (18)	52*(18)	22 (15)
February	38 (42)	41*(16)	9 (17)	-10 (22)	24 (21)	-41*(18)	15 (18)
March	1 (41)	-2 (15)	-11 (13)	-18 (21)	29 (21)	6 (19)	-4 (14)
April	55 (40)	-2 (13)	17 (17)	17 (21)	36 (20)	0 (17)	-13 (16)
May	70 (44)	20 (15)	17 (13)	12 (27)	14 (21)	4 (20)	3 (15)
June	110*(41)	41*(15)	10 (14)	6 (21)	35 (18)	13 (20)	5 (17)

6 Accounting for discontinuities in the time series model

The parallel run showed that the redesign resulted in discontinuities in the series of the monthly figures about the labour force. To avoid severe model misspecification, the intervention term $\mathbf{\Delta}_t\boldsymbol{\beta}$ has to be included in model (3.1). An additional question is how the available information about the discontinuities in the first panel, obtained with the parallel run, can be used efficiently in the time series model. Six different methods to use the available information from the parallel run in model (3.1) and (3.9) are discussed.

Method 1: Model (3.1) with a diffuse prior for all intervention variables.

The time independent regression coefficients of the intervention variables for all five panels are included in the state vector and initialised with a diffuse prior, as described by Durbin and Koopman (2001), Subsection 6.2.2. The Kalman filter can be applied straightforwardly to obtain estimates for the regression coefficients. This approach ignores the information about the discontinuities that is available from the parallel run. In this application, this approach is interesting since comparing the time series model estimate for the discontinuity in the first panel with the direct estimates obtained with the parallel run illustrates how well discontinuities can be estimated with the intervention approach.

Method 2: Model (3.1) with an exact prior for the intervention variable of the first panel.

The direct estimates of the discontinuities from the parallel run are incorporated into the model by using an informative prior for the initialization of β^1. This can be done by using these estimates in the initial state vector for β^1 and their estimated variances as an uncertainty measure for β^1 in the covariance matrix of the initial state vector.

Method 3: Model (3.1) where the regression coefficient of the intervention variable for the first panel equals the average direct estimate for the discontinuity obtained with the parallel run.

Another possibility of using the direct estimate of the discontinuities in the first panel as a-priori information in model (3.1), is to assume that the regression coefficient for the intervention in the first panel is time independent and equal to the average value of the observed discontinuity in the parallel run, i.e.,

$$\bar{\beta}^1 = \frac{1}{6}\sum_{t=\tilde{t}}^{\tilde{t}+5}\left(\hat{Y}_t^{t,\text{New}} - \hat{Y}_t^{t,\text{Old}}\right),$$

where $\tilde{t}$ denotes the start of the parallel run in January 2010. In this case the direct estimate for the discontinuity is treated as if it is a fixed value, known in advance. This approach ignores the uncertainty of using a survey estimate for the discontinuity.

Method 4: As method 3, but with a time dependent regression coefficient for the intervention variable of the first panel.

The direct estimates for the discontinuities fluctuate considerably over the six months of the parallel run, see Table 5.2. To have a smooth transition from the old to the new design, an alternative for method 3 is considered where during the parallel run, the regression coefficient of the first panel is time dependent and equals the observed monthly discontinuities. For the period after the parallel run, this regression coefficient is equal to the average value of the observed discontinuity in the parallel run, i.e.,

$$\beta_t^1 = \begin{cases} \hat{Y}_t^{t,\text{New}} - \hat{Y}_t^{t,\text{Old}} & \text{if} \quad t \in [\tilde{t},\dots,\tilde{t}+5] \\ \bar{\beta}^1 & \text{if} \quad t > \tilde{t}+5. \end{cases}$$

This method comes down to replacing the observations under the new design by the observations under the old design during the parallel run and assumes that the results under the old design are more reliable during this period. Similar to method 3, the uncertainty of using a survey estimate for the discontinuity is ignored.

The four methods can be applied to model (3.9) that is extended with an auxiliary series about the number of people formally registered at the employment office. The following two methods are considered:

Method 5: Equals Method 1 applied, to model (3.9).

Method 6: Equals Method 4 applied, to model (3.9).

In practise, method 1 would be considered if no parallel run is available. In the case of a well conducted parallel run, method 2 is probably the most natural approach, because the sample estimate for the discontinuity together with its uncertainty are used as prior information in the model. The sample information that becomes available after the parallel run under the new design is still used to improve the estimate of the discontinuity. Methods 3 and 4 are considered as alternatives for method 2 for getting a smoother transition from the estimates obtained until June 2010 under the old design to the estimates under the new design, starting in July 2010. Method 3 might work well if the variation between the monthly estimates for the discontinuity during the parallel run is small. In the case of large fluctuations between the monthly discontinuities, method 4 might be considered because during the parallel run each monthly deviation of the estimate under the new design is nullified with the time dependent discontinuities. Method 4 will therefore result in the smoothest transition.

In the case of strong and reliable auxiliary information, each method can be combined with model (3.9). It is a requirement, however, that the evolution of this auxiliary series is not influenced by factors that are unrelated to the real developments of the labour market. Method 5 would be considered if no parallel run is available. The auxiliary series might result in more precise estimates for the discontinuity and the trend and signal of the unemployed labour force. In the case of a parallel run, method 2 in combination with model (3.9) is probably the most natural approach for similar reasons as mentioned before (results not presented). Method 6 can be used to get a smoother transition from the old to the new design and more precise estimates for the trend and the signal of the unemployed labour force by taking advantage of the available auxiliary information. For similar reasons method 3 can be combined with model (3.9) (results not presented).

7 Results

7.1 Estimation results for the national level

Results are presented for the monthly unemployed labour force figures at the national level. The filtered estimates for the discontinuities in panels 1 and 2 are plotted in Figure 7.1. Figure 7.2 compares the filtered estimates of the RGB in panel 2 under the six different methods from January 2006 until March 2012, and the filtered RGB obtained under the old data until June 2010. Results for the other panels are similar and therefore omitted. Figure 7.3 compares the filtered trend estimates under the six different methods from July 2009 until March 2012, with the filtered trend estimates obtained under the old data until June 2010.

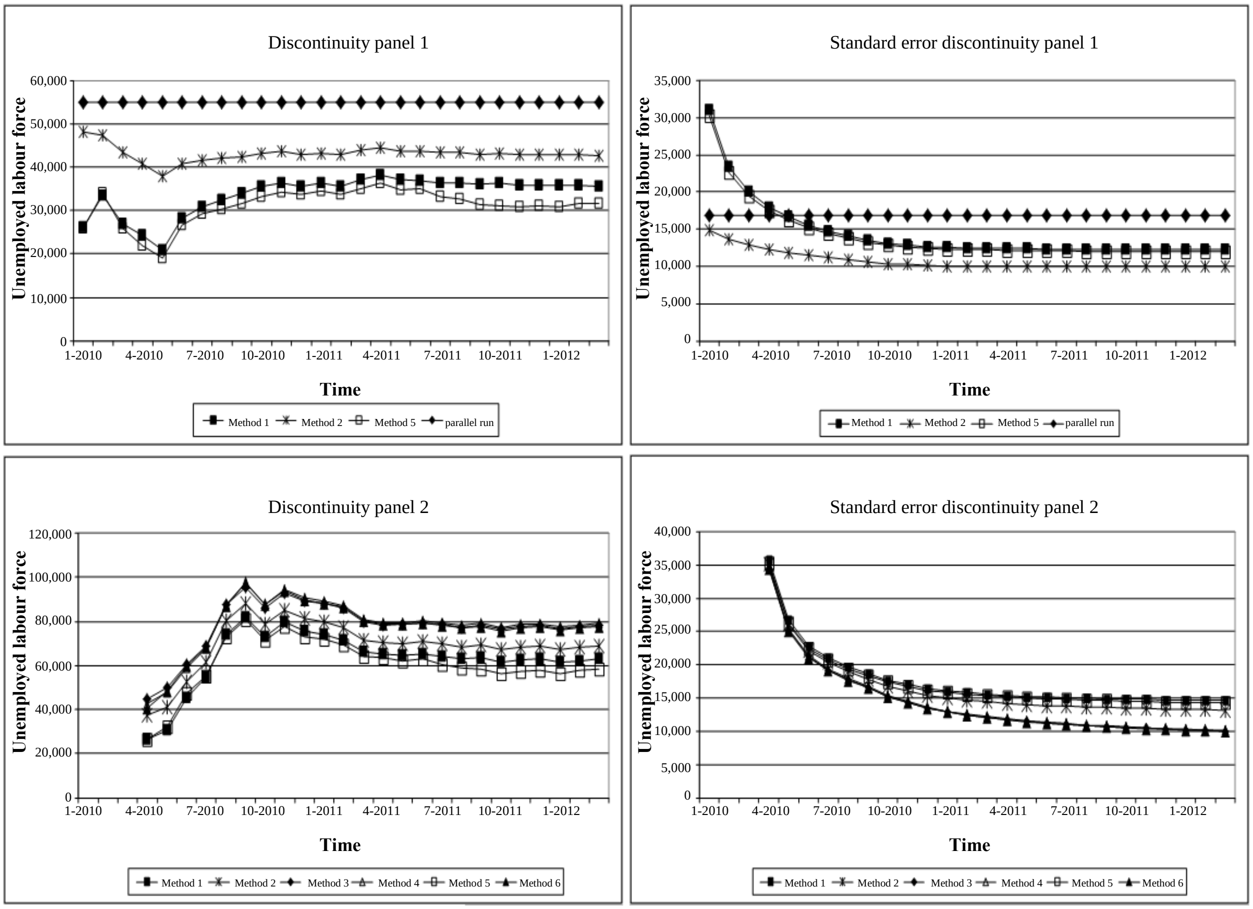
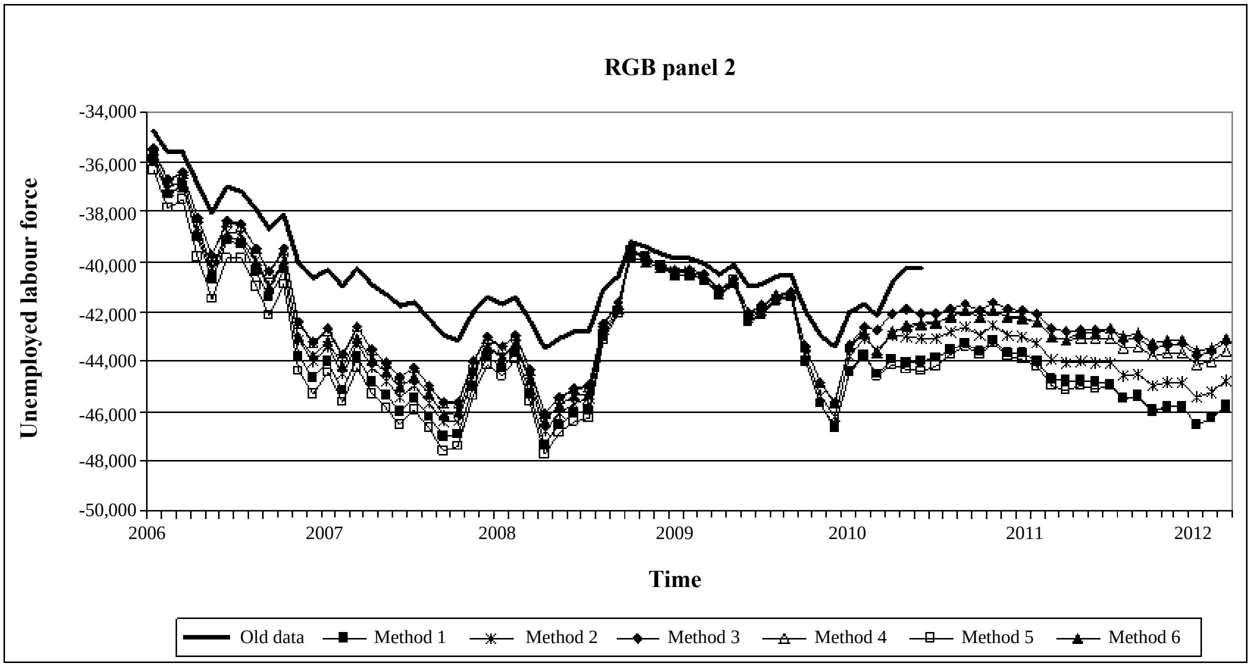


Figure 7.1 Filtered estimates for discontinuities and their standard errors January 2010 – March 2012, panel 1 and 2 for monthly unemployed labour force at national level.



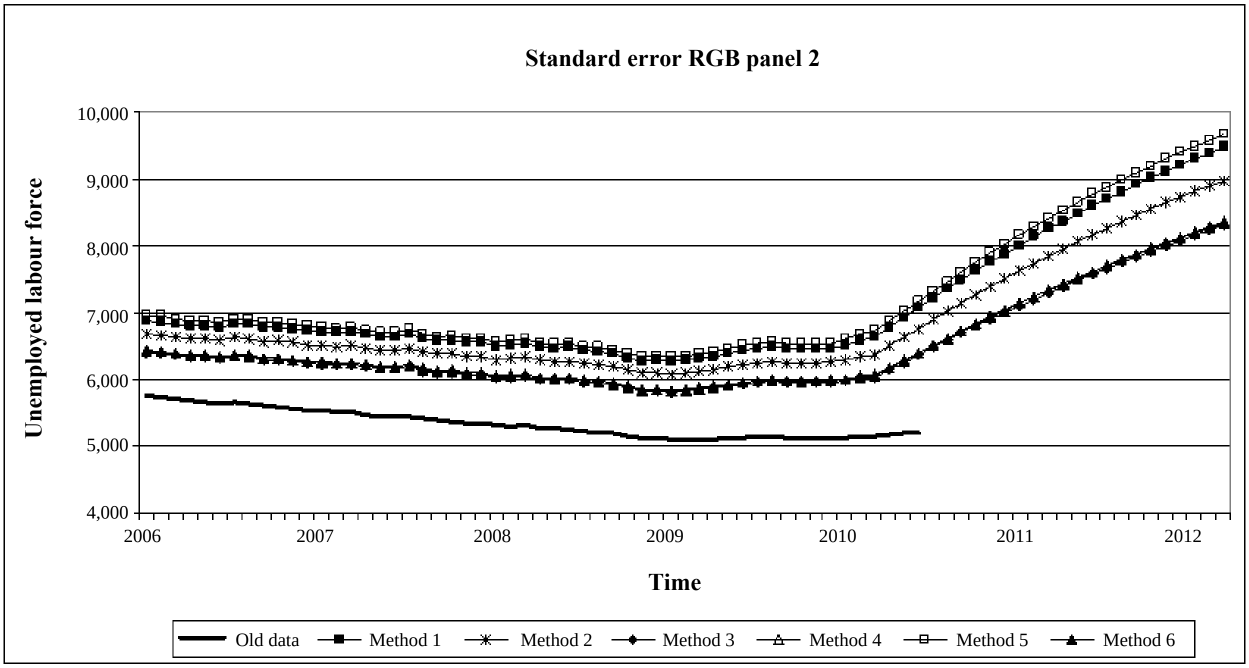


Figure 7.2 Filtered estimates for RGB and their standard errors panel 2 for monthly unemployed labour force at national level January 2006 – March 2012 for six different methods that account for discontinuities and the old data.

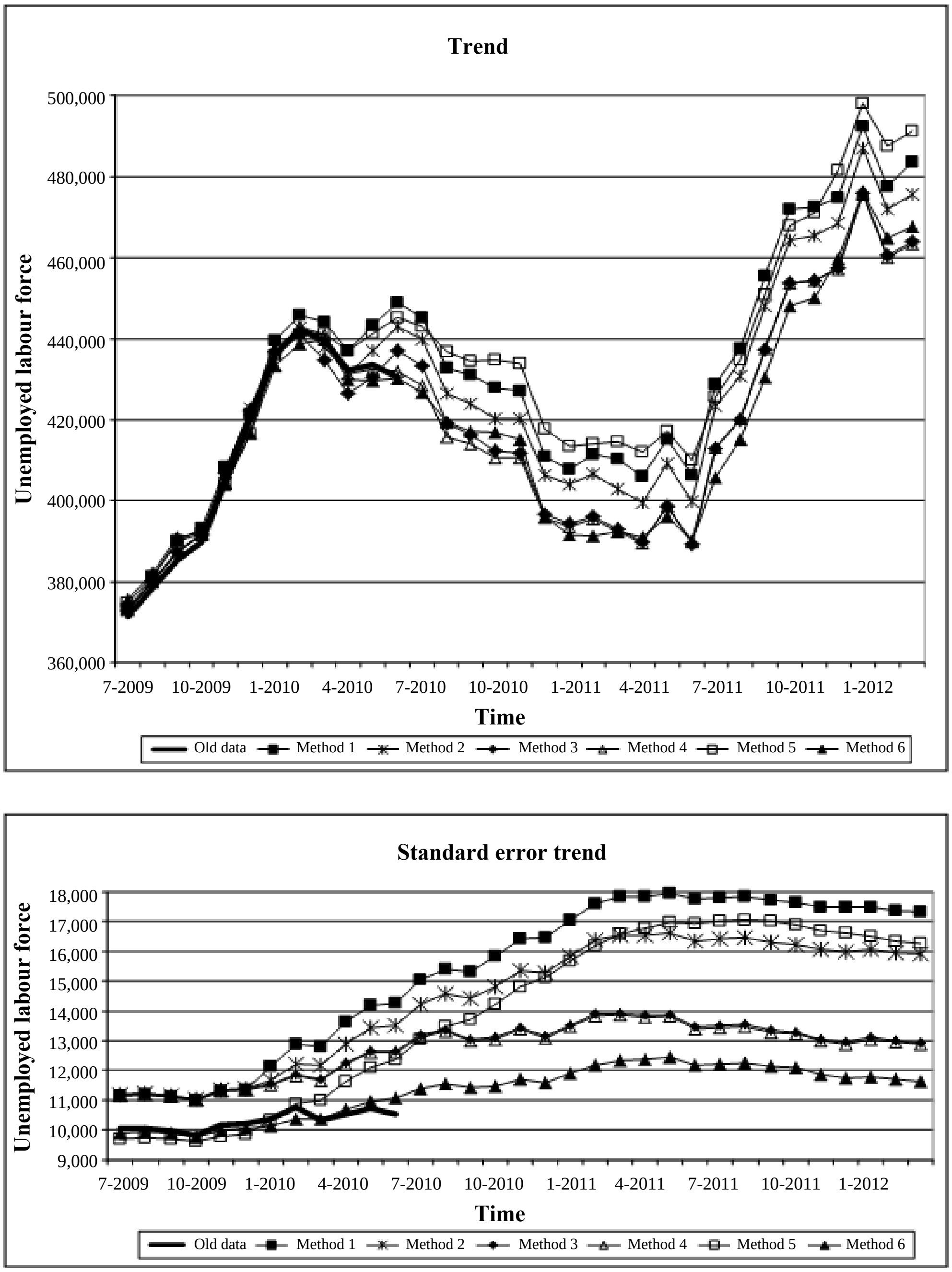


Figure 7.3 **Filtered trend of monthly unemployed labour force at the national level and their standard errors July 2009 - March 2012 for six different methods that account for discontinuities and the old data.**

Figure 7.1 shows that the different methods lead to different estimates for the discontinuities. The filtered estimates for the regression coefficient of the intervention variable in the first panel are systematically smaller than the direct estimate obtained in the parallel run. The smallest estimate is obtained if a diffuse prior is used to initialise this regression coefficient (method 1 and 5). Extending the model with an auxiliary series resulted in a slightly smaller estimate (compare method 1 and 5). Using the direct estimate from the parallel run as an exact prior for the regression coefficient (i.e., method 2) resulted, as expected, in an estimate that is closer to the direct estimate obtained with the parallel run.

The standard errors of the regression coefficients of the interventions follow a smooth exponentially decreasing pattern. Already five months after the change-over to the new design, the standard errors of the regression coefficients initialised with a diffuse prior became smaller than the standard error of the direct estimate for the discontinuity obtained in the parallel run. The standard errors of the regression coefficients initialised with an exact prior were, as expected, immediately smaller than the standard error of the direct estimate.

The estimated discontinuities in panel 2 through 5 follow the same pattern as the estimates observed in panel 1. Methods with small estimates for the discontinuity in panel 1, also have the smallest estimates in the subsequent panels and vice versa. As described below, the estimate of the discontinuity in the first panel strongly influences the estimated level of the trend. This explains why the method used to quantify the discontinuity in the first panel also influences the estimated discontinuities in the subsequent panels. Extending the model with an auxiliary series hardly affects the estimated discontinuities (method 6 versus 3 and 4, method 5 versus 2). On average the estimated regression coefficients become more or less stable about one year after the change-over. By using the exact prior in the first panel (method 2), a stable estimate for the discontinuity in the first panel is obtained after about half a year. The auxiliary series, on the other hand, do not decrease the required period to obtain a stable estimate.

The filtered estimates for the discontinuities are affected by the model choice of the RGB. Since the model for the RGB is time dependent, the filtered estimates for the RGB may partially absorb the discontinuities induced by the redesign. Therefore the filtered estimates for the regression coefficients do not reflect the absolute effect of the redesign. They nevertheless avoid model misspecification due to discontinuities in the input series. More realistic estimates for the discontinuities are obtained with a model were the RGB is time invariant $(\text{i.e., } \sigma_\lambda = 0)$. Under this model, the estimated discontinuities for the first panel indeed increase with about 7,000 persons under method 1, 2, and 5 and come closer to the direct estimate for the discontinuity observed in the parallel run (results not presented).

The standard errors of the regression coefficients in panel 2 through 5 are affected by the method used to estimate the discontinuity in the first panel. Method 3, 4 and 6, which use the direct estimate from the parallel run for the discontinuity in the first panel have the smallest standard errors and are more or less equal. Method 1 and 5, which use a diffuse prior for the regression coefficient for the discontinuity in the first panel, have the largest standard errors for the discontinuities in the subsequent panels. Method 2, which uses an exact prior in the first panel, has standard errors that are somewhere in between.

Figure 7.2 shows that the filtered RGB is also influenced by the intervention term and the method used to estimate the discontinuity in the first panel. Most striking is the difference between the RGB with the data observed under the old approach only, and the RGB obtained with the six methods that include the data under the new approach, during the period before the change-over to the new design. These differences can be explained with differences between the ML estimates for the hyperparameter of the RGB $(\hat{\sigma}_\lambda)$. Adding the data observed under the new design and augmenting the model with an appropriate intervention term increases $\hat{\sigma}_\lambda$ with a factor of about 1.4 (compare Tables 4.1 and 7.2).

After the change-over to the new design, the estimates for the RGB become less volatile than in the period before the change-over. The level of the RGB after the change-over also depends on the method used to quantify the discontinuity in the first panel. As will be explained below, the value for the discontinuity in the first panel determines the level of the trend in the first panel and therefore also the relative bias, i.e., the RGB, in the subsequent panels with respect to the first panel.

The evolution of the standard errors of the filtered RGB shows a smooth pattern. The standard errors for the RGB under the old design are substantially smaller since the ML estimate for the hyperparameter is smaller compared to the methods that include the data observed under the new design. The introduction of the five intervention variables, starting in January 2010, introduced additional uncertainty in the estimated RGB. As a result the standard errors consistently increased after January 2010. It is remarkable that they did not stabilize within the observed period, like the standard errors of the trends (see below). This might be caused by the fact that the discontinuities simultaneously influence the intervention variables and the RGB parameters and could be an indication that the model has difficulties separating both effects with a model that allows for time dependent RGB. A model with constant RGB has indeed a constant standard error for the RGB after the change-over. The problems with model identification increased with the flexibility of the RGB.

The order of the standard errors of the RGB under the six methods is equal to the results observed for the standard error of the discontinuities. Similar results hold for the RGB in the other three panels.

Figure 7.3 shows that the level of the trend (and also the signal) strongly depends on the choice of the method used to estimate the discontinuities. Larger estimates for the discontinuities resulted in smaller levels for the trend and vice versa. The evolution of the trend is more or less similar under the six methods.

Before the change-over, the standard errors of the trend under the new design were larger compared to the method that only uses the old data, with the exception of method 5 and 6, which are based on the model extended with an auxiliary variable. This difference can be attributed to the increased flexibility of the RGB as described before. Methods 5 and 6 have more or less the same standard error as the method based on the old data only. The disturbance terms of the slope of the auxiliary series and the monthly unemployed labour force were strongly correlated (about 0.9). This resulted in a substantial decrease of the standard error of the filtered trend and neutralized the increase of the standard error due to the increased flexibility of the RGB.

Each time a panel changes to the new design, the standard error of the filtered trend increases under each of the six methods and stabilizes after the change-over in the fifth panel. Methods 1 and 5, which use a diffuse prior for the discontinuity in the first panel showed the largest increases in the standard error at each time a new intervention variable modelled the change-over to the new design in a panel. Recall from Figure 7.1 that the standard errors for the discontinuities in the five panels are the largest under these two methods. The standard error for the trend under method 5 is smaller than in method 1, since this method takes advantage of a strongly correlated auxiliary series. Method 2, which uses an exact prior, follows more or less the same pattern, but had smaller increases in the standard error. Methods 3, 4, and 6, which use the direct estimate for the discontinuity in the first panel, had the smallest increase in the standard error of the trend, since they had the smallest standard error for the four discontinuities in panel 2 through 5 and ignored the standard error of the direct estimate for the discontinuity in the first panel. The standard errors for method 3 and 4 were equal. The standard errors for method 6 were smaller since this method benefited from the correlated auxiliary series.

We do not present the results for filtered slopes and seasonals but just mention that the standard errors of these state variables are not affected by the change-over to the new design in the different panels.

7.2 Estimation results for domains

Roughly speaking, similar results are observed for the six domains. Table 7.1 summarizes the trend and the discontinuities in the first panel with their standard errors averaged over the last 12 months of the six domains and the national level for the six methods. For method 5 and 6 the ML estimates for the correlation between the disturbances of the slopes are also included. The differences between the direct estimates for the discontinuities and the regression coefficients of the intervention in the first panel are in some cases larger compared to the national level. This can be expected since the sample size in the domains is smaller, resulting in less precise direct estimates for the discontinuities.

Methods 5 and 6, which take advantage of a correlated auxiliary series, showed a stronger decrease for some of the domains of the standard error of the filtered trend compared to the national level. In these cases, the ML estimates of the correlation were larger and sometimes equal to one, which implies that the trend of the auxiliary series and the unemployed labour force are or tend to be cointegrated.

Table 7.1
Trend and discontinuities panel 1 averaged over the last 12 months of the national level and the six domains for the six different methods used to quantify the discontinuity in the first panel. Standard errors between brackets

Parameter	Method	National level	Men 15-24	Women 15-24	Men 25-44	Women 25-44	Men 45-64	Women 45-64
Trend	1	452 (18)	58 (5)	40 (5)	78 (8)	100 (7)	87 (7)	82 (6)
	2	445 (16)	53 (5)	41 (5)	83 (8)	95 (7)	85 (7)	79 (6)
	3	435 (13)	45 (4)	44 (4)	95 (6)	83 (5)	82 (5)	73 (4)
	4	434 (13)	45 (4)	44 (4)	95 (6)	83 (5)	82 (5)	73 (4)
	5	454 (17)	58 (4)	43 (4)	78 (8)	98 (6)	77 (4)	83 (6)
	6	433 (12)	45 (4)	45 (3)	92 (5)	83 (3)	76 (3)	74 (4)
Disc. panel 1	1	36 (12)	5 (4)	11 (4)	17 (6)	3 (5)	2 (5)	-4 (5)
	2	43 (10)	11 (3)	10 (3)	11 (5)	8 (4)	3 (4)	-2 (4)
	5	33 (12)	6 (3)	10 (4)	15 (6)	5 (5)	5 (4)	-5 (5)
	3, 4, 6	55 (17)	19 (6)	7 (6)	-1 (9)	20 (8)	6 (8)	4 (7)
Corr. slope	5	0.93	0.98	0.99	0.93	0.99	1.00	0.87
	6	0.88	0.72	1.00	0.99	1.00	1.00	0.90

7.3 Model choice

As a consequence of the strong correlation between the disturbances of the slopes, the auxiliary series has a notable effect on the level of the filtered trend. Using a model that includes this auxiliary series therefore implies that there must be great confidence in the quality of the auxiliary series. Amendments in the law with respect to unemployment benefits and social benefits, or sudden changes in the mode of operation of the employment office, may result in sudden or gradual differences in the number of people formally registered at the employment office. This would not be a problem if the ML estimates for the correlation between the disturbances of the slopes became smaller. Simulations where level breaks as well as gradual increasing disturbances are added to the auxiliary series show that the ML estimates for the correlation are adjusted with an unacceptable large delay. Therefore the auxiliary series may influence the filtered trend estimates for the monthly unemployed labour force incorrectly (results not shown). Since it is known that the evolution of the series of the number of people formally registered at the employment office is influenced by the aforementioned factors, that are unrelated to economic developments, it was decided not to choose methods 5 or 6 to produce official monthly unemployment figures.

The model diagnostics, mentioned in the second paragraph of Section 4, indicate that the innovations under model (3.9) contain more autocorrelation and slightly stronger deviations from the normality assumption than model (3.1). The model diagnostics for the four methods based on model (3.1) are very similar and do not indicate strong violations of the assumption that the innovations are normally and independently distributed. The model diagnostics are not useful for further discriminating between the different methods that rely on the same model (model (3.1) or (3.9)). This is a consequence of the interchange between the estimates for the discontinuities and the trend. As explained before, an increase in the estimated discontinuity is neutralized by an opposite effect on the filtered trend and RGB. As a result, the one-step-a-head predictions for the signals and the innovations in the different panels are more or less equal under all methods.

The main purpose of modelling discontinuities is to avoid that developments of labour force indicators are erroneously influenced by the change-over to the new survey process. The preferred method describes the development of the monthly labour force figures most accurately. The choice between methods 1 through 4 can therefore be based on the confidence in the different estimates for the discontinuities, using additional information such as knowledge from subject matter experts. Comparing the filtered trends under the different methods with the officially published figures during the parallel run is also useful for evaluating which method results in the smoothest transition during the change-over.

Recall from Section 5 that the model estimates obtained under the old data were published as the official monthly release until June 2010. The figure to be published for July 2010 must be based on one of the new methods, where the observations in the time series of the first panel changed from the old to the new method in January 2010 (see Section 5). From Figure 7.3 it follows that during the parallel run the filtered trend obtained with method 4 is, from the methods based on model (3.1), the closest to the officially published trend obtained with the old data. It can therefore be expected that this method will result in the smoothest transition in the month that the data under the new approach are used for the first time. According to labour market experts, there were no indications that the steady downward trend of the monthly unemployed labour force could change into an upward trend at that time. As follows from Figure 7.3, method 4 is the only method based on model (3.1) that resulted in a continued downward trend.

Based on the aforementioned considerations, method 4 was finally chosen to produce official statistics about the monthly labour force. With method 4, the GREG estimates in the first panel were corrected back to the outcomes under the old design during the parallel run, and this resulted in the smoothest and most plausible transition to the new method.

7.4 Implementation

ML estimates for the hyperparameters based on method 4 at the national level and the six domains are presented in Table 7.2. In Figure 7.4, the five GREG series are plotted with the filtered trend based on the model, which is currently used to produce official model-based estimates for the monthly unemployed labor force figures. The detail of this figure is not important. The purpose is to illustrate how noisy the five input series of the GREG estimates are and how, with the time series model, a filtered trend from this input is obtained. Until 2010, the level of the filtered trend was equal to the level of the GREG estimates of the first panel, since the model removes the RGB by benchmarking the outcomes to the level of the series obtained in the first panel. In 2010 the change-over to the new design started. The discontinuities resulted in higher levels for the series of GREG estimates of the five panels. In this application, the time series model estimates figures that are corrected for these discontinuities. As a result, the filtered trend drops below the level of the series observed with the first panel after 2010.

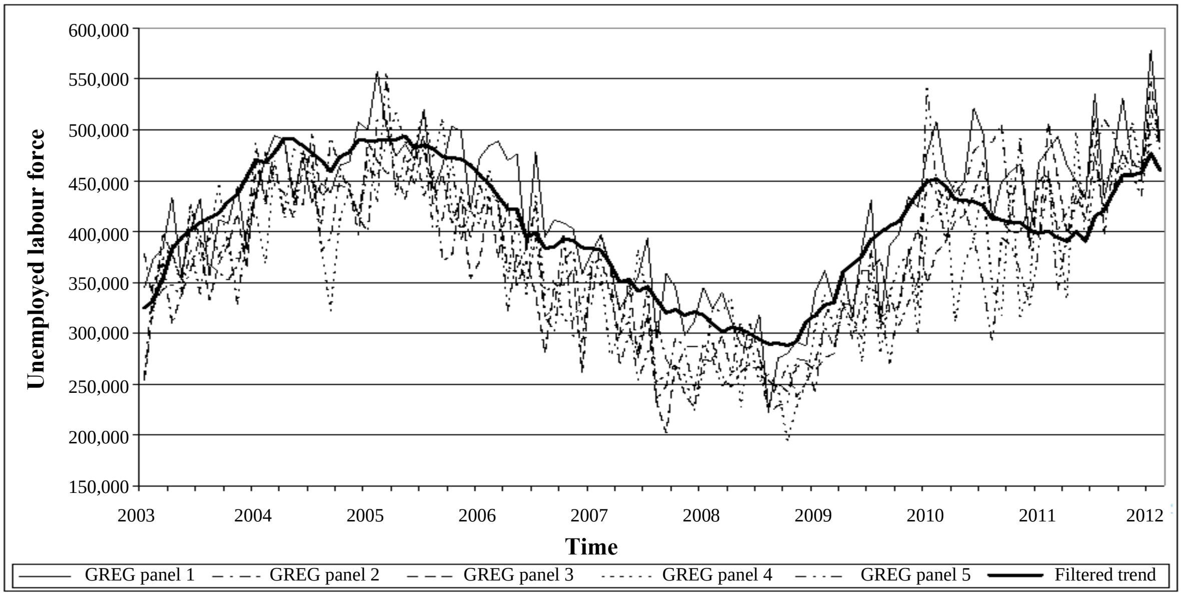


Figure 7.4 **Unemployed labour force at the national level; GREG estimates of the five panels and filtered trend based on a structural time series model.**

Table 7.2
ML estimates of hyperparameters for monthly unemployed labour force figures after the survey redesign. Values are expressed as standard deviations

Standard deviation		National	Men	Women	Men	Women	Men	Women
		level	15-24	15-24	25-44	25-44	45-64	45-64
Slope	$(\hat{\sigma}_\eta)$	2,423	292	221	703	561	451	207
Seasonal	$(\hat{\sigma}_\omega)$	0.01	0.09	0.00	0.00	0.00	0.00	0.00
RGB	$(\hat{\sigma}_\lambda)$	1,218	931	654	316	567	272	418
White noise	$(\hat{\sigma}_\varepsilon)$	7,720	1,663	3,348	4,128	4,540	4,383	3
Survey error panel 1	$(\hat{\sigma}_{e1})$	0.99	0.93	1.02	1.03	0.97	0.99	1.13
Survey error panel 2	$(\hat{\sigma}_{e2})$	1.03	0.95	1.10	1.16	1.00	1.18	1.14
Survey error panel 3	$(\hat{\sigma}_{e3})$	0.96	1.05	1.15	1.15	1.00	1.18	1.00
Survey error panel 4	$(\hat{\sigma}_{e4})$	1.12	1.05	1.17	1.16	1.03	1.13	1.07
Survey error panel 5	$(\hat{\sigma}_{e5})$	1.13	1.02	1.08	1.11	1.04	1.17	1.01
Autocorrelation	$(\hat{\rho})$	0.257	0.130	0.212	0.430	0.245	0.456	0.411

The hyperparameter estimates for the survey errors for panel 2, 3, 4 and 5 are divided by $(1-\hat{\rho}^2)$. Therefore hyperparameters for the survey errors are, as expected, around 1.

The filtered estimates, considered so far, illustrate what can be accomplished with the state-space approach to produce contemporary estimates in the production of official statistics, i.e., the optimal estimates for period t based on the sample information observed until period t. These filtered estimates, however, can be improved if new information after period t becomes available. Although Statistics Netherlands currently does not revise the contemporary estimates, it is interesting to analyze to what extent the filtered estimates are adjusted if information that becomes available after one, two or three months is used to update the filtered estimates. In Figure 7.5 the filtered trend $\left(L_{t|t}\right)$ is compared with the estimates based on the information of one $\left(L_{t|t+1}\right)$, two $\left(L_{t|t+2}\right)$ and three $\left(L_{t|t+3}\right)$ additional months after period t for the unemployed labour force at the national level. The smoothed series based on the entire series is also included in this figure.

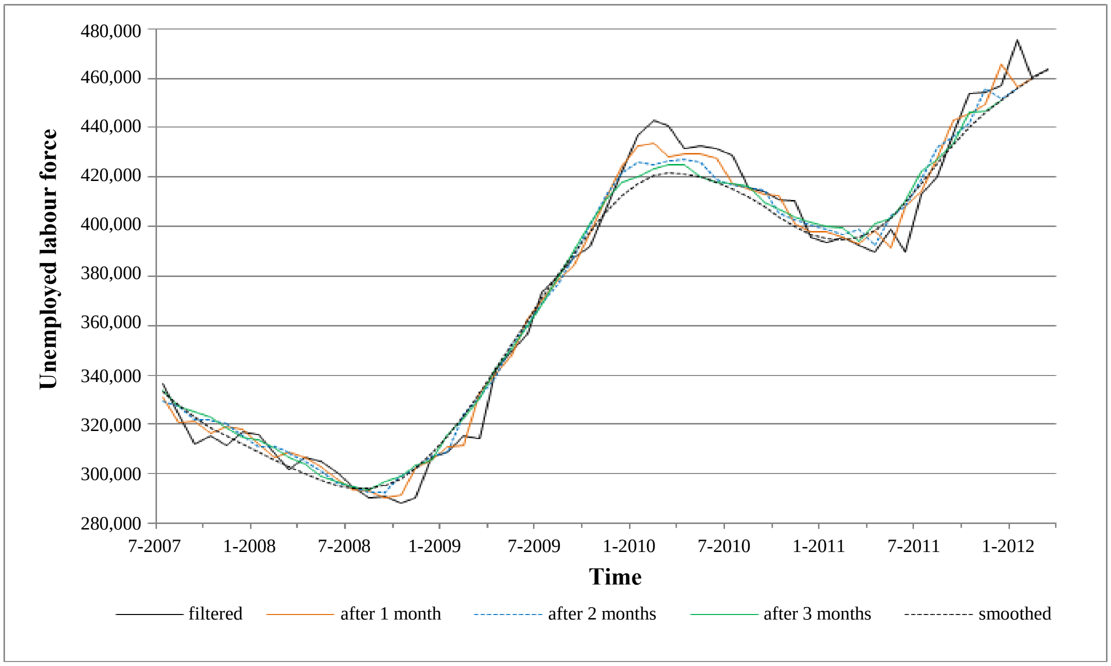


Figure 7.5 **Comparison of filtered trend, revisions after one month $\left(L_{t|t+1}\right)$, two months $\left(L_{t|t+2}\right)$, three months $\left(L_{t|t+3}\right)$, and the smooth trend for the unemployed labour force at the national level.**

The largest revisions occur if the information after the first three months are used to update the filtered estimates. The estimates based on the information observed after three months are already close to the smoothed estimates. Furthermore the revisions during the period of the change-over, starting in January 2010, are larger than in other periods. This is the result of the introduction of the intervention variables. Particularly in this period, the estimates for the intervention variables are based on a few observations under the new design, resulting in large revisions for the discontinuities. This is reflected in larger revisions for the trend during the period of the change-over. In this application, it appears that the first two or three month after period t contain substantial additional information to improve the monthly estimate for period t. It could therefore be considered to base the final estimates for period t on the information observed until $t + 2$ or $t + 3$.

8 Discussion

National statistical institutes widely apply GREG estimators to produce official statistics. The advantage of these estimators is that they are robust against model misspecification, reduce the design variance, and correct at least partially for selection bias in the case of well-specified weighting models. Furthermore, they result in domain estimates which are consistent by definition, and their use in production processes is relatively straightforward since only one set of weights is required to estimate all possible output tables in a multipurpose survey.

GREG estimators, however, have unacceptably large design variances in the case of small sample sizes and do not handle measurement bias in an effective way. The Dutch LFS is an example where these problems require additional estimation procedures. The sample size is too small to produce sufficiently precise monthly labour force figures with the GREG estimator. The rotating panel design and the major redesign of the survey process make differences in measurement bias visible and compromises comparability of outcomes over time. These problems are solved simultaneously with a multivariate structural time series model that uses the series with GREG estimates for the different panels as input. The time series method combines strong points of the GREG estimator with the advantages of a model-based approach. Since time series of GREG estimates as well as their standard errors are used as input series, the method accounts for the complexity of the sample design and corrects for unequal selection probabilities and selective non-response. The time series model accounts for small sample sizes by taking advantage of sample information observed in previous periods, the autocorrelation in the survey errors, the rotation group bias by benchmarking the estimates to the level of the first panel, and discontinuities that arise from a major survey redesign.

We discussed how the model can be extended with a strongly correlated auxiliary series, which is the number of people formally registered at the employment office in this application. Auxiliary information further decreases the standard error of the filtered trend and signal. Also the levels of the filtered estimates are affected by the auxiliary variable. Since there are strong indications that the evolution of the auxiliary series is affected by factors other than economic cycles, and that this improperly affects the monthly filtered trend of the unemployed labour force, it was decided not to use this information in the ultimately selected model. In this application, the auxiliary series hardly influences the estimated discontinuities. This conclusion, however, cannot be generalized. If e.g., the moment of the change-over coincides with a real break in the evolution of the variable of interest, then auxiliary series should contain similar breaks and can provide valuable additional information to disentangle discontinuities from real developments correctly.

If no parallel run is conducted, then discontinuities are estimated through an intervention variable with a regression coefficient initialized with a diffuse prior. In the case of a parallel run, direct estimates for the discontinuities provide additional information that can be used in the time series model. One possibility is to use the direct estimate with its standard error as an exact prior to initialize the regression coefficient of the intervention variable. Another approach is to assume that the regression coefficient is equal to the direct estimate. This approach treats the external information about the discontinuities as if it is observed without error. A well-conducted parallel run has the advantage that it provides a direct estimate for the discontinuities and therefore does not rely on the assumption that, at the moment of the change-over, the evolution of the variables of interest is captured by the time series components other than the intervention variable.

A consequence of modelling discontinuities is that the standard errors of the filtered trend and signal increase each time the new design enters another panel. This illustrates the importance of keeping the survey process unchanged as long as possible and of limiting the number of redesigns.

In conclusion, a time series model is proposed that simultaneously solves problems with small sample sizes, RGB in a rotating panel, and discontinuities due to a redesign. It enables Statistics Netherlands to publish real monthly figures about the labour force, instead of the rolling quarterly figures that are often used as a second best approximation. During the redesign, the model avoids distortion of real developments of the monthly labour force indicators with sudden changes in measurement bias. The method is flexible and of general interest, since most national statistical institutes apply rotating panels for labour force surveys. Furthermore, redesigns of survey processes aimed to reduce administration costs or to improve outdated methods remain inevitable, resulting in loss of comparability of the outcomes over time. Finally there is an increasing interest for small area estimates while there is always pressure to reduce sample sizes due to budget constraints and lowering the response burden.

Acknowledgements

The authors wish to thank the referees and the Associate Editor and Rita Gircour (Statistics Netherlands), for careful reading of the first draft of this paper and providing constructive comments. The views expressed in this paper are those of the authors and do not necessarily reflect the policy of Statistics Netherlands.

Appendix

With the structural time series model (3.1), monthly estimates for the employed, unemployed and the total labour force are computed for the national level and for a breakdown in the six domains. These 21 population parameters are notated by $\theta_{t,l,m}$, where $l = 1, 2, 3$ denotes respectively the employed, unemployed and total labour force, $m = 1$ the national level, and $m = 2, \ldots, 7$ the six domains. For the population parameters, the following consistency requirements hold:

$$\theta_{t,1,m} + \theta_{t,2,m} - \theta_{t,3,m} = 0, m = 1, \ldots, 7 \tag{A.1}$$

$$\sum_{m=2}^{7} \theta_{t,l,m} = \theta_{t,l,1}, l = 1, 2, 3. \tag{A.2}$$

Subscript m runs within l, which in turn runs within t. Because time series model (3.1) is applied to each population parameter separately, requirements (A.1) and (A.2) do not hold for the model estimates.

Therefore, they are restored with a Lagrange function. The model estimates for the national level are changed as little as possible, because they are based on considerably larger samples than the six domains. Therefore, the consistency is achieved in two steps. Both steps are specified for the filtered trends. Consistent filtered signals can be computed in a similar way.

Let $L_{t,l,m}$ denote the filtered trend for $\theta_{t,l,m}$. In the first step, the requirements of equation (A.1) for the national level $(m = 1)$ are considered. The consistency requirement can be written as $\mathbf{\Delta}^{[1]}\mathbf{L}_t^{[1]} = 0$ with $\mathbf{L}_t^{[1]} = \left(L_{t,1,1}, L_{t,2,1}, L_{t,3,1}\right)^T$ a vector with the model estimates for the three trends at the national level and $\mathbf{\Delta}^{[1]} = (1, 1, -1)$ a 3×1 matrix that specifies requirement (A.1). Adjusted estimates that fulfil (A.1) are computed with the Lagrange function

$$\mathbf{L}_{t,\text{adj}}^{[1]} = \mathbf{L}_t^{[1]} - \mathbf{V}_t^{[1]}\mathbf{\Delta}^{[1]T}\left(\mathbf{\Delta}^{[1]}\mathbf{V}_t^{[1]}\mathbf{\Delta}^{[1]T}\right)^{-1}\mathbf{\Delta}^{[1]}\mathbf{L}_t^{[1]} \tag{A.3}$$

with $\mathbf{L}_{t,\text{adj}}^{[1]} = \left(L_{t,1,1,\text{adj}}, L_{t,2,1,\text{adj}}, L_{t,3,1,\text{adj}}\right)^T$ the adjusted filtered trends. In the ideal case $\mathbf{V}_t^{[1]}$ is the variance-covariance matrix of the trend estimates $\mathbf{L}_t^{[1]}$. The covariances of the model estimates, however, are not known. Therefore the diagonal matrix of the variances is used instead.

In the second step, $\mathbf{L}_{t,\text{adj}}^{[1]}$ is not changed anymore. Now the vector of domain estimates $\mathbf{L}_t^{[2]} = \left(L_{t,1,2}, L_{t,1,3}, \ldots, L_{t,1,7}, L_{t,2,2}, \ldots, L_{t,2,7}, L_{t,3,2}, \ldots, L_{t,3,7}\right)^T$ is adjusted according to equation (A.1) for $m = 2, \ldots, 7$ and to equation (A.2) for $l = 1, 2$. Equation (A.2) for $l = 3$ is redundant and therefore left out. Again, the consistency requirements for the filtered trends of the domains are written as $\mathbf{\Delta}^{[2]}\mathbf{L}_t^{[2]} = \mathbf{C}_t^{[2]}$, with

$$\mathbf{\Delta}^{[2]} = \begin{pmatrix} \mathbf{I}_6 & \mathbf{I}_6 & -\mathbf{I}_6 \\ \mathbf{1}_6 & \mathbf{0}_6 & \mathbf{0}_6 \\ \mathbf{0}_6 & \mathbf{1}_6 & \mathbf{0}_6 \end{pmatrix},$$

$\mathbf{C}_t^{[2]} = \left(\mathbf{0}_6, L_{t,1,1,\text{adj}}, L_{t,2,1,\text{adj}}\right)^T$, $\mathbf{I}_6$ the six dimensional identity matrix, and $\mathbf{1}_6$ and $\mathbf{0}_6$ six dimensional row vectors with each element equal to one or zero respectively. Consistent domain estimates are computed with the Lagrange function

$$\mathbf{L}_{t,\text{adj}}^{[2]} = \mathbf{L}_t^{[2]} - \mathbf{V}_t^{[2]}\mathbf{\Delta}^{[2]T}\left(\mathbf{\Delta}^{[2]}, \mathbf{V}_t^{[2]}, \mathbf{\Delta}^{[2]T}\right)^{-1}\left(\mathbf{\Delta}^{[2]}\mathbf{L}_t^{[2]} - \mathbf{C}_t^{[2]}\right),$$

similarly to (A.3). In this case $\mathbf{V}_t^{[2]}$ is the diagonal matrix of the variances of the estimates of $\mathbf{L}_t^{[2]}$.

References

Bailar, B.A. (1975). The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, 70, 23-30.

Binder, D.A., and Dick, J.P. (1990). A method for the analysis of seasonal ARIMA models. *Survey Methodology*, 16, 2, 239-253.

Doornik, J.A. (2009). An Object-oriented Matrix Programming Language Ox 6. London: Timberlake Consultants Press.

Durbin, J., and Koopman, S.J. (2001). *Time Series Analysis by State Space Methods*. Oxford: Oxford University Press.

Harvey, A.C., and Durbin, J. (1986). The effects of seat belt legislation on British road casualties: A case study in structural time series modelling. *Journal of the Royal Statistical Society, Series A,* 149, 187-227.

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

Koopman, S.J. (1997). Exact initial Kalman filtering and smoothing for non-stationary time series models. *Journal of the American Statistical Association*, 92, 1630-1638.

Koopman, S.J., Shephard, N. and Doornik, J.A. (2008). *SsfPack 3.0: Statistical Algorithms for Models in State Space Form*. London: Timberlake Consultants Press.

Pfeffermann, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business & Economic Statistics*, 9, 163-175.

Pfeffermann, D., Feder, M. and Signorelli, D. (1998). Estimation of autocorrelations of survey errors with application to trend estimation in small areas. *Journal of Business & Economic Statistics*, 16, 339-348.

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

Robinson, P.M., and Särndal, C.-E. (1983). Asymptotic properties of the generalized regression estimator in probability sampling. *Sankhyā B*, 45, 240-248.

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

U.S. Census Bureau (2009). X-12-ARIMA Reference Manual. Washington DC.

van den Brakel, J.A., and Krieg, S. (2009). Estimation of the monthly unemployment rate through structural time series modelling in a rotating panel design. *Survey Methodology*, 35, 2, 177-190.

van den Brakel, J.A., and Roels, J. (2010). Intervention analysis with state-space models to estimate discontinuities due to a survey redesign. *Annals of Applied Statistics*, 4, 1105-1138.

Survey Methodology, December 2015
Vol. 41, No. 2, pp. 297-314
Statistics Canada, Catalogue No. 12-001-X

Domain sample allocation within primary sampling units in designing domain-level equal probability selection methods

Avinash C. Singh and Rachel M. Harter[1]

Abstract

Self-weighting estimation through equal probability selection methods (*epsem*) is desirable for variance efficiency. Traditionally, the *epsem* property for (one phase) two stage designs for estimating population-level parameters is realized by using each primary sampling unit (PSU) population count as the measure of size for PSU selection along with equal sample size allocation per PSU under simple random sampling (SRS) of elementary units. However, when self-weighting estimates are desired for parameters corresponding to multiple domains under a pre-specified sample allocation to domains, Folsom, Potter and Williams (1987) showed that a *composite measure of size* can be used to select PSUs to obtain *epsem* designs when besides domain-level PSU counts (i.e., distribution of domain population over PSUs), frame-level domain identifiers for elementary units are also assumed to be available. The term *depsem*-A will be used to denote such (one phase) two stage designs to obtain domain-level *epsem* estimation. Folsom et al. also considered two phase two stage designs when domain-level PSU counts are unknown, but whole PSU counts are known. For these designs (to be termed *depsem*-B) with PSUs selected proportional to the usual size measure (i.e., the total PSU count) at the first stage, all elementary units within each selected PSU are first screened for classification into domains in the first phase of data collection before SRS selection at the second stage. Domain-stratified samples are then selected within PSUs with suitably chosen domain sampling rates such that the desired domain sample sizes are achieved and the resulting design is self-weighting. In this paper, we first present a simple justification of composite measures of size for the *depsem*-A design and of the domain sampling rates for the *depsem*-B design. Then, for *depsem*-A and -B designs, we propose generalizations, first to cases where frame-level domain identifiers for elementary units are not available and domain-level PSU counts are only approximately known from alternative sources, and second to cases where PSU size measures are pre-specified based on other practical and desirable considerations of over- and under-sampling of certain domains. We also present a further generalization in the presence of subsampling of elementary units and nonresponse within selected PSUs at the first phase before selecting phase two elementary units from domains within each selected PSU. This final generalization of *depsem*-B is illustrated for an area sample of housing units.

Key Words: *Epsem* and *depsem* designs; Multiple domain estimation; Self-weighting estimation; Two phase two stage designs.

1 Introduction

For multi-stage design of surveys, an equal probability selection method (or *epsem*, Kish 1965, page 21) is typically desired toward the goal of variance reduction or variance efficiency. In practice, for two or more stage designs, selection probabilities for primary (or first stage) sampling units (PSUs) are often driven by considerations of over- (under-) sampling to obtain adequate domain sample sizes, and operational efficiency such as equal interviewer workload per PSU. The simplest type of an *epsem* design is a single stage simple random sampling (SRS) design without replacement of elementary units with selection probabilities n/N where n, N denote respectively the sample and population sizes. Another example is single stage stratified SRS with proportional allocation; i.e., $n_h / N_h \propto 1$, or $n_h = fN_h$ where $f = n/N$, and n_h, N_h are sample and population sizes, respectively, for the h^{th} stratum. These and other *epsem* designs are described in fundamental sampling texts such as those by Cochran (1977) and Lohr (2010).

[1] Avinash C. Singh, NORC at the University of Chicago, 55 East Monroe St., 20th Floor, Chicago, IL 60603. E-mail: singh-avi@norc.org; Rachel M. Harter, RTI International, P.O. Box 12194, Research Triangle Park, NC 27709.

Yet another example of an *epsem* design is single stage SRS of whole clusters. For area sampling in field surveys, clusters are useful for operational efficiency due to reduced travel cost in interviewing neighboring housing units although there are some drawbacks. Cluster sizes could vary considerably making the logistics difficult for equalizing interviewer workloads. Moreover, a complete enumeration of each cluster may not be desirable due to cost, and inefficient estimation due to reduced effective sample size as a result of intra-cluster correlations. In general, the probability proportional to size (pps) sampling of clusters followed by equal sample allocation of elementary units per cluster to equalize interviewer assignments is a reasonable and practical compromise for area cluster sampling.

Above considerations lead to two stage designs with first stage selection probabilities to be denoted by π_i for the i^{th} PSU, and second stage conditional selection probabilities to be denoted by $\pi_{j|i}$ for the j^{th} elementary unit within the i^{th} selected PSU. For example, in a survey of teachers, PSUs could be schools, while ultimate sampling or elementary units could be teachers within schools. For SRS of size n_i^* within each PSU i with population count N_i, the probabilities $\pi_{j|i}$ and π_i can be defined as follows to obtain an *epsem* design; see Kish (1965, page 222). Here n_i^* are common and equal to n/m where m is the desired number of selected PSUs out of a total of M PSUs in the population. We have

$$\pi_i = m\frac{N_i}{N},\ \pi_{j|i} = \frac{n_i^*}{N_i} = \frac{n}{mN_i}. \tag{1.1}$$

It is easily seen, as expected, that the sum of π_i's over all M PSUs i is the fixed sample size m at the first stage, and the sum of $\pi_{j|i}$'s over all N_i elementary units j within the i^{th} PSU is the fixed sample size n_i^* at the second stage. Moreover, the unconditional (same as joint because of nesting of units within PSUs) selection probability for the j^{th} unit in the i^{th} PSU is the product $\pi_i\pi_{j|i}$; i.e., n/N or f, which is equal for all units, as desired. For generalizations of self-weighting estimation considered in this paper, it is useful to express the implied sample allocation n_i to the i^{th} PSU from (1.1) as

$$n_i = \left(\frac{f}{\pi_i}\right)N_i, \tag{1.2}$$

based on the observation $\pi_i \times n_i/N_i$ equals f where f is the desired sampling fraction n/N. Here, the value of n_i is obtained as $n_i^*\,(= n/m)$. Note that if all PSUs are selected with certainty; i.e., $\pi_i = 1$, the above PSU$-$level allocation reduces to proportional allocation in stratified designs with the number of strata being the total number M of PSUs.

The basic idea for making any design *epsem* is to work backwards; that is, before specifying selection probabilities π_i for PSUs, it is ensured that the sampling rate within any given PSU i is inversely proportional to π_i so that π_i cancels out in the unconditional selection probability $\pi_i\pi_{j|i}$ within the PSU. In this way the unconditional selection probabilities for elementary units can be made common for all sampled units from different PSUs. We will use this strategy throughout the paper.

From (1.1), observe that in order for $n_i \le N_i$, we must have $f \le \pi_i$ for all $i = 1,\ldots,M$. This condition can be satisfied at the design stage by collapsing neighboring PSUs in order to increase N_i (and hence π_i) or by reducing f if necessary. In other words, the sample allocated to the i^{th} PSU must be a fraction of the PSU population size N_i where the fraction is given by the ratio of the desired sampling rate f and the PSU selection probability π_i.

So far we considered *epsem* designs for a single domain; i.e., estimation at the population level only. However, often survey designs are intended to support analytical goals for multiple domains within the target population. For example, in the case of a teacher survey, domains could be male and female teachers. For domain-level *epsem* designs (to be termed *depsem* in this paper), Folsom, Potter and Williams (1987) presented a method for allocating a sample of units to PSUs under two separate designs - $depsem-\text{A1}$ and $depsem-\text{B1}$ defined as follows; the numeric extension in the notation is used to differentiate them from other variations presented later.

The $depsem-\text{A1}$ *Design* is defined as a one phase two stage design where domain-level PSU population counts (N_{id} for the i^{th} PSU and d^{th} domain), desired domain sample size $\left(n_{+d}^*\right)$ where '+' denotes sum over m selected PSUs, and equal PSU sample allocation $n_{i+}^*\ (= n/m)$ over all domains (i.e., equal interviewer load) are specified. Thus, the desired sampling rate (f_d) for each domain is pre-specified but the PSU selection probabilities (π_i) are not pre-specified and are suitably defined to obtain the *depsem* property. Here it is also assumed that frame-level domain identifiers for elementary units are available. Such a design is applicable to situations where in-person interviews with a list frame are desirable.

The $depsem-\text{B1}$ *Design* is defined as a two phase two stage design where PSU population counts (N_{i+}) and desired domain sample size $\left(n_{+d}^*\right)$ over m selected PSUs are specified. Domain-level population counts (N_{+d}) are not specified (which of course implies that domain-level PSU population counts N_{id} are not specified), and $\text{PSU}-\text{level}$ sample allocations (n_{i+} over all domains) are also not pre-specified. In addition, the desired sampling rates for each domain (f_d) are not pre-specified. However, PSU selection probabilities are specified by using PSU population counts as size measures, and for selected PSUs in the first stage, domain-level population counts N_{id} become available after the first phase census. Here the domain sampling rates f_d are suitably defined to obtain the *depsem* property. The two phase aspect of the design is used to obtain domain membership of selected units in the first phase through screening. Such a design may be applicable more generally than the previous one.

The school/teacher example can be used to make these two *depsem* designs concrete. In $depsem-\text{A1}$, we know in advance how many male and female teachers are in each school from the list frame, and also we know which teachers are male and which are female. The desired sampling rates of male and female teachers, and the equal number of teachers to be selected per school are known. Then school or PSU selection probabilities are obtained to satisfy the *depsem* property. In $depsem-\text{B1}$, we know the probability of selecting each school based on the total number of teachers per school. We do not know how many male and female teachers are in each school, but the desired numbers of male and female teachers in the sample over all selected schools are specified. Then, after screening all teachers in the selected schools for male/female classification, the sampling rates for male and female domains for each pre-selected school are obtained to satisfy the *depsem* property.

For $depsem-\text{A1}$, Folsom et al. (1987) provide a composite measure of size for selecting PSUs such that its inverse appears in the specification of domain sample allocations within each PSU. The sample allocation to domains within PSUs satisfies the desired PSU sample size or interviewer workload exactly. However, the desired domain sample size is achieved only in expectation because the sample size of elementary units within domains is not directly controlled, but the PSU sample size is controlled instead to obtain equal interviewer workload.

For $depsem - \mathrm{B1},$ the same basic method is inverted to produce $depsem$ samples. Here, in the first phase, a census of selected PSUs at the first stage is conducted so that all elementary units within selected PSUs are stratified into domains to obtain domain-level PSU counts and are subsampled such that the desired domain sample size over all PSUs is satisfied. However, any constraint on the PSU sample size is relaxed in the interest of obtaining a $depsem$ sample. $Depsem - \mathrm{B1}$ may be particularly useful for non-face-to-face interview modes such as telephone surveys in the second phase, where the first phase sample of elementary units is used to obtain contact information and domain classification. The first phase results may be based on a self-administered screening questionnaire sent by mail or dropped off after an in-person contact effort to all or a large sample of units in each selected PSU. If the main interview is conducted by phone in the second phase, having equal interviewer workload per PSU is of no practical consequence. Folsom et al. (1987) also considered natural generalizations of both $depsem$ designs to the case of stratified population of PSUs in the first phase.

In this paper, we introduce a systematic general framework for defining $depsem$ designs which provides a simple justification for the $depsem$ property of the above two designs. We then propose generalizations of the two designs under the above framework to obtain new useful variations of $depsem$ designs encountered in practice; see Singh and Harter (2011) for an earlier development. See also Fahimi and Judkins (1991) for an interesting simulation study comparing traditional and nontraditional measures of size with respect to between PSU variance contributions. The organization of this paper is as follows. Section 2 reviews the original composite measure of size method for selecting PSUs as proposed by Folsom et al. (1987) for the $depsem - \mathrm{A1}$ design including its stratified version. Section 3 presents the inverted method of Folsom et al. (1987) for $depsem - \mathrm{B1}$ to obtain domain-level sampling rates over all pre-selected PSUs. Section 4 presents a generalization to a hybrid $depsem - \mathrm{AB}$ design where the domain-level PSU counts for all PSUs are assumed to be only approximately known, and are used first to specify PSU selection probabilities obtained as composite measures of size as in $depsem - \mathrm{A1},$ and then sampling rates from selected PSUs are specified as in $depsem - \mathrm{B1}$ by obtaining true domain-level PSU counts for selected PSUs through first phase screening. Another generalization considered in Section 4 is when PSUs in the first phase are selected with arbitrarily pre-specified selection probabilities. Section 5 further generalizes $depsem - \mathrm{B1}$ to designs where the second phase sample within each selected PSU is not a census (i.e., there is subsampling within PSUs) or when it is a census but is subject to nonresponse, or both. Generalizations to stratified designs are also considered in Section 5. Section 6 presents a hypothetical but realistic example based on a study for which the proposed $depsem$ designs were developed under a two-phase two stage design to establish nationally representative norms for an English and Spanish instrument toolbox for assessing behavioral and cognitive functions. We conclude with remarks in Section 7.

2 Review of $depsem - \mathrm{A}$ designs with a simple justification

Consider a one phase two-stage design where the first stage units are schools, for example, and the second stage units are individual teachers. Two domains of interest may be male and female teachers. Under $depsem - \mathrm{A1},$ it is assumed that the $\mathrm{PSU} - \text{domain}$ population counts (N_{id}) are known, where the

PSU index i varies from 1 to $M,$ the total number of schools; and the domain index d varies from 1 to $D,$ where D in this example is 2 for male and female teachers. In addition, it is assumed that the frame-level domain identifiers (male/female) are available for each teacher in the list. Now, suppose the desired number of sampled teachers for each domain d is n_{+d}^{*} based on precision requirements, where the subscript '+' in n_{+d}^{*} denotes aggregation over selected PSUs i varying from 1 to $m.$ The sum of n_{+d}^{*} over all domains is the total sample size $n.$ Then we know the desired sampling rate for domain d teachers is $f_d = n_{+d}^{*} / N_{+d}$ where N_{+d} is the sum of N_{id} for domain d across all M schools. In addition, it is desired to have equal sample sizes in all m selected schools; i.e., $n_{i+}^{*} = n/m$ for $i = 1, \ldots, m.$

Folsom et al. (1987) proposed a composite measure of size for selecting schools which can be used to allocate the desired number of sampled teachers within schools in such a way that the selected teachers provide *epsem* designs for both male and female teacher domains. The design satisfies exactly the specified equal sample size $n_{i+}^{*} \, (= n/m)$ for all selected schools but only in expectation the specified domain sample size $n_{+d}^{*}.$ Clearly, it is practical to control directly the sample size within each selected school and not the domain sample size overall selected schools.

We provide a different but simpler derivation of the results given in Folsom et al. (1987). To this end, we observe that the key result (1.2) for *epsem* designs implies that the sampling rate n_{id} / N_{id} in domain d within PSU i should be proportional to $f_d / \pi_i.$ This is true regardless of how the PSU selection probabilities π_i or the domain sampling rates f_d are specified. For $depsem-$A1, although frame-level domain identifiers for elementary units are assumed to be known, it may not be cost efficient to directly draw samples from domains after stratifying the frame. It may be preferable to select PSUs in the first stage which are then stratified by domains using frame-level information before the second stage sample selection using SRS. So in the interest of equal interviewer workload per PSU, we consider the allocation of the desired sample size n_{i+}^{*} for a given PSU i to domains so that the PSU i sample size is controlled at the desired value. However, the realized domain sample size then becomes random and can be made to satisfy the desired goal in expectation.

***Depsem*$-$A1 Design:** For each given PSU $i,$ the sample allocations n_{id}^{A1} over domains are obtained as $n_{id}^{A1} \propto f_d N_{id} / \pi_i,$ which implies that

$$n_{id}^{A1} = \left(n_{i+}^{*}\right) \frac{f_d N_{id} / \pi_i}{\sum_{d'=1}^{D} f_{d'} N_{id'} \Big/ \pi_i} = \left(\frac{n}{m}\right) \frac{f_d N_{id}}{S_i}, \tag{2.1}$$

where S_i denotes $\sum_{d'=1}^{D} f_{d'} N_{id'}$ as the unspecified π_i cancels out. However, we can set $\pi_i^{A1},$ the selection probability for PSU $i,$ as $mS_i / S_+,$ where $S_+ = \sum_{i=1}^{M} \sum_{d=1}^{D} f_d N_{id}.$ By exchanging summations, S_+ reduces to $\sum_{d=1}^{D} f_d \sum_{i=1}^{M} N_{id}$ or $\sum_{d=1}^{D} f_d N_{+d} \, (= n).$ Then the allocated sample size n_{id}^{A1} over domains can be expressed analogous to (1.2) as

$$n_{id}^{A1} = \left(\frac{f_d}{\pi_i^{A1}}\right) N_{id} \; . \tag{2.2}$$

Observe that if $\pi_i^{A1} = 1$; i.e., if PSUs are selected with certainty, the above allocation behaves like the proportional allocation in stratified designs for domains within PSUs acting as strata. It is easy to show from equation (2.2) that the probability of an individual teacher j being selected is equal for all sampled teachers in domain d where teachers are selected by a stratified SRS from each selected PSU stratified by domains. The probability depends only on d because

$$\text{Pr (teacher } j \text{ in domain } d \,|\, \text{school } i\text{) Pr (school } i\text{)} = \frac{n_{id}^{A1}}{N_{id}} \pi_i^{A1} = f_d. \tag{2.3}$$

Thus S_i, a composite measure of size, provides the appropriate size measure for PSU i to obtain a $depsem-$A1 design. Unlike the traditional size measure given by the PSU population count N_i used in population level $epsem$ designs, the new size measure S_i depends on the desired domain sample size n_{+d}^* as well as the domain-level PSU population size N_{id} because of domain-level $epsem$ requirements. The measure S_i can be interpreted as the approximate total desired sample size over all domains within each PSU i.

It is also observed that for PSU i while the sample allocations $\{n_{id}^{A1}\}_{1\le d\le D}$ over domains satisfy the desired sample size n_{i+}^* exactly by construction $\left(\text{i.e.,} \sum_{d=1}^{D} n_{id}^{A1} = n_{i+}^*\right)$, the resulting allocations $\{n_{id}^{A1}\}_{1\le i\le m}$ for any given domain d over selected PSUs satisfy the desired sample size n_{+d}^* only in expectation; i.e.,

$$E\left(n_{+d}^{A1}\right) = f_d E\left(\sum_{i=1}^{m} \frac{N_{id}}{\pi_i^{A1}}\right) = \frac{n_{+d}^*}{N_{+d}} E\left(\hat{N}_{+d}^{A1}\right) = n_{+d}^*, \tag{2.4}$$

where $\hat{N}_{+d}^{A1}$ denotes $\sum_{i=1}^{m} N_{id} \big/ \pi_i^{A}$ and estimates N_{+d} unbiasedly, and E is the expectation operator for the first stage randomization.

It should be remarked that in practice the allocations $\{n_{id}^{A1}\}_{1\le d\le D}$ need not be integers, and may require random rounding. To do this, consider the fractional parts $\{n_{id}^{A1} - [n_{id}^{A1}]\}_{1\le d\le D}$ where [.] denotes the greatest integer contained in the quantity in brackets. These fractional parts in a PSU can be treated as selection probabilities for selecting without replacement a sample of size defined by the sum of the fractional parts for that PSU, which is necessarily an integer. Then allocations for domains so selected are rounded up, while for others they are rounded down. Thus the randomly rounded domain allocations continue to satisfy the condition of fixed sample size n_{i+}^*, but the desired domain allocation n_{+d}^* is now satisfied under the joint expectation of random design and rounding mechanisms.

The above derivation of sample allocation assumed implicitly that $n_{id}^{A1} \le N_{id}$; i.e., the allocated sample size does not exceed the corresponding population size. This assumption requires that the factor $\left(f_d \big/ \pi_i^{A1}\right)$ must be less than or equal to 1 for all d and i in view of (2.2). In other words, we must have

$$\max\{f_d\}_{1\le d\le D} \le \min\{\pi_i^{A1}\}_{1\le i\le M}. \tag{2.5}$$

By reducing values of f_d, or by collapsing neighboring PSUs to increase π_i^{A1}, it is generally not difficult in practice to satisfy the above condition. Incidentally, randomly rounded n_{id}^{A1}'s continue to satisfy (2.5) if the original n_{id}^{A1}'s do.

***Depsem*$-$A2 Design:** All the above results easily generalize to stratified two stage designs denoted by $depsem-\text{A2}$; e.g., schools may be stratified by school districts in our simple example. Specifically, the key result (2.2) is generalized to obtain the domain sample allocations $\{n_{hid}^{A2}\}_{1\le d\le D}$ of n_{hi+}^{*} within PSU i of the h^{th} stratum, $h=1,\ldots,H$, as follows:

$$n_{hid}^{A2}=\left(\frac{f_d}{\pi_{hi}^{A2}}\right)N_{hid}, \tag{2.6}$$

where notations with subscript h signify that the terms are stratum-specific. Other results mentioned above for the unstratified case can be easily extended in an analogous manner to the stratified case.

3 Review of *depsem*$-$B designs with a simple justification

Now suppose that the schools have already been sampled with usual PSU population counts N_i as size measures, and therefore their probabilities of selection π_i are known as given in (1.1). Under $depsem-\text{B1}$ involving two phase designs, f_d's are not specified, but the desired values of n_{+d}^{*} are pre-specified for all domains. For example, the schools are pre-selected and the desired numbers of sampled male and female teachers are pre-specified, but the sampling rates for male and female teachers are not specified. It is still possible to select *epsem* samples of male and female teachers using suitable values of f_d in (2.2), but not with equal sample size per school, as shown in Folsom et al. (1987) and described below.

***Depsem*$-$B1 Design:** Here π_i is set by the usual size measure as mN_i/N for $i=1,\ldots,M$. Denote it by π_i^{B1}. As in $depsem-\text{A1}$, the sampling rate n_{id}/N_{id} in domain d within PSU i should be set proportional to f_d/π_i^{B1}, although here f_d is not known. Under $depsem-\text{B1}$, each selected PSU is stratified by domains for the selection of elementary units using the first phase domain-screening information while under $depsem-\text{A1}$, domain memberships of elementary units are assumed to be available in the frame itself. In this case, the condition of equal interviewer workload per PSU is relaxed and the desired PSU sample sizes n_{i+}^{*} over all domains are not pre-specified. Instead, it is the desired domain sample size n_{+d}^{*} that is directly controlled by allocating it to PSUs within each domain d. Thus, n_{+d}^{*} is rendered nonrandom which is clearly preferable for control on resulting precision of domain level estimates. It follows that, analogous to (2.1), the sample allocations $\{n_{id}^{B1}\}_{1\le i\le m}$ of the domain total n_{+d}^{*} for each domain d to selected PSUs are given by

$$n_{id}^{B1}=\left(n_{+d}^{*}\right)\frac{f_d N_{id}/\pi_i^{B1}}{\sum_{i'=1}^{m} f_d N_{i'd}/\pi_{i'}^{B1}}=\left(\frac{n_{+d}^{*}}{\hat{N}_{+d}^{B1}}\right)\frac{N_{id}}{\pi_i^{B1}}=\left(\frac{\hat{f}_d^{B1}}{\pi_i^{B1}}\right)N_{id}, \tag{3.1}$$

where $\hat{N}_{+d}^{B1}=\sum_{i'=1}^{m}N_{i'd}/\pi_{i'}^{B1}$ and the unspecified f_d cancels out. However, we can set $\hat{f}_d^{B1}$, the sampling rate for domain d, as $n_{+d}^{*}/\hat{N}_{+d}^{B1}$. Clearly, $\{n_{id}^{B1}\}_{1\le i\le m}$ satisfies n_{+d}^{*} exactly by construction. However, the allocations do not satisfy n_{i+}^{*} (or n/m), even in expectation, because in general

$$E\left(n_{i+}^{B1}\right) = E\left(\frac{\sum_{d=1}^{D}\hat{f}_d^{B1}N_{id}}{\pi_i^{B1}}\right) = \left(\frac{n}{m}\right)\frac{\sum_{d=1}^{D}E\left(\hat{f}_d^{B1}\right)N_{id}}{\sum_{d=1}^{D}fN_{id}} \neq \frac{n}{m}, \tag{3.2}$$

unless $\hat{f}_d^{B1}$ is constant and equals $f\,(= n/N)$, which is in conflict with the desired disproportionate domain allocations.

Other considerations such as random rounding of n_{id}^{B1} to obtain integer allocations carry over in a manner analogous to $depsem-\mathrm{A1}$. However, if the requirement of $n_{id}^{B1} \leq N_{id}$ for all domains within each $i = 1, \ldots, m$, is not satisfied, one option is to reduce n_{+d}^{*}, while the other option is to collapse neighboring PSUs. For example, collapsing i and i', and letting $\tilde{i}$ denote the collapsed PSU, we have $N_{\tilde{i}d} = N_{id} + N_{i'd}$. Then $\pi_{\tilde{i}}^{B1}$ required for calculating sample allocations in the second phase from (3.2) is now given by $\pi_{\tilde{i}}^{B1} = \pi_i^{B1} + \pi_{i'}^{B1} - \pi_{ii'}^{B1}$ which, incidentally, also requires knowledge of the second order inclusion probability $\pi_{ii'}^{B1}$.

***Depsem*–B2 Design:** We next consider a generalization of the above case to stratified designs. In our example of the teacher survey, this case corresponds to schools stratified by school districts. This extension carries over in a manner analogous to $depsem$ – A2. That is, suppose for the first phase sample, m_h PSUs are to be selected from the h^{th} stratum, $h = 1, \ldots, H$ with the usual pre-specified selection probabilities $m_h N_{hi} / N_h$ to be denoted by π_{hi}^{B2}. The sample allocations $\left\{n_{hid}^{B2}\right\}_{1\leq i\leq m_h, 1\leq h\leq H}$ of the domain total n_{++d}^{*} to selected PSUs within each stratum h, analogous to formula (3.1) of $depsem-\mathrm{B1}$, are given by

$$n_{hid}^{B2} = \left(\frac{\hat{f}_d^{B2}}{\pi_{hi}^{B2}}\right)N_{hid}, \tag{3.3}$$

where N_{hid} is the domain d population count within PSU i and stratum h,

$$\hat{f}_d^{B2} = n_{++d}^{*} \Big/ \hat{N}_{++d}^{B2}, \text{ and } \hat{N}_{++d}^{B2} = \sum_{h=1}^{H}\sum_{i=1}^{m_h} N_{hid} \Big/ \pi_{hi}^{B2}.$$

4 Proposed generalizations of *depsem* – A/B designs

$Depsem-\mathrm{A1}$ and $-B1$ designs require relatively stringent assumptions regarding the provision of frame-level domain membership information of elementary units for $depsem-\mathrm{A}$ designs and domain-screening of all elementary units from selected PSUs in the first phase for $depsem-\mathrm{B}$ designs. In practice, the assumptions may not be true exactly, yet the goal of $depsem$ sample designs may still be desirable. In this section we loosen the requirement for $depsem-\mathrm{A1}$ that domain-level PSU counts are known exactly which leads to a new hybrid design $depsem-\mathrm{AB1}$ where PSU – domain counts are initially assumed to be only approximately known in order to specify PSU selection probabilities as in $depsem-\mathrm{A1}$. Later, true domain-level PSU counts for selected PSUs at the first stage are obtained as in $depsem-\mathrm{B1}$ by conducting a census of elementary units within PSUs in the first phase. Another design termed $depsem-\mathrm{C}$ generalizes $depsem-\mathrm{B}$ by employing a general pre-specification of PSU selection probabilities. Both cases use the same strategy of making the domain-level PSU sampling rates inversely proportional to the PSU selection probabilities. Table 4.1 provides a quick summary of old and new designs considered in this paper.

Table 4.1
Summary of different *depsem* designs (old and new)

***Depsem* design description unstratified (or stratified)**	**PSU selection probability π_i (or π_{hi})**	**Domain-level PSU population count N_{id} (or N_{hid})**	**Domain sampling rate f_d**	**Domain sample size n_{+d} (or n_{++d})**	**PSU sample size n_{i+} (or n_{hi+})**
A1 (or A2): One phase two stage (Old)	Find	Specified (also frame-level domain identifiers)	Specified	Specified (in expectation)	Specified
B1(or B2): Two phase two stage (Old)	Specified	Obtain from phase one census of selected PSUs	*Find*	Specified	Not specified
AB1 (or AB2): Hybrid one/two phase two stage (New)	Specified using A1 and initial values $\tilde{N}_{id}$	Specified approximate initial values $\tilde{N}_{id}$ for all PSUs; and exact values N_{id} for selected PSUs from phase one census	*Find*	Specified	Not specified
C1 (or C2): Two phase two stage (New)	Specified	Specified from first phase census of selected PSUs	*Find*	Specified	Not specified
$C1'$ (or $C2'$): Two phase two stage with subsampling and nonresponse at phase one (New)	Specified (also response and subsampling rates within PSUs)	Specified for selected PSUs from phase one respondents	*Find*	Specified	Not specified

***Depsem*–AB1 Design:** Consider a new case for *depsem* designs using a variation of $depsem-\mathrm{A1}$ in which domain-level PSU population counts N_{id} are only approximately known and given by $\tilde{N}_{id}$. The approximations may be available from alternative sources such as the most recent census or a suitable administrative database. In our teacher example, the number of male and female teachers in each school may be known for the prior year, which serves as an approximation for the current year domain-level PSU counts. The m PSUs are selected using π_i^{AB1} probabilities which, similar to π_i^{A1} under $depsem-\mathrm{A1}$, are defined as

$$\pi_i^{AB1} = m\sum_{d=1}^{D} \tilde{f}_d \tilde{N}_{id} \Big/ n,\ \tilde{f}_d = n_{+d}^{*} \big/ \tilde{N}_{+d}\,. \tag{4.1}$$

Now we consider two phases in addition to two stages, as in $depsem-\mathrm{B1}$, because first stage units within selected PSUs need to be classified into domains, and corresponding true counts N_{id}'s are to be determined. In this case, all elementary units in the PSU are selected in the first phase sample. Now, analogous to formula (3.1) of $depsem-\mathrm{B1}$, the sample allocations $\{n_{id}^{AB1}\}_{1\le i\le m}$ of the domain sample size n_{+d}^{*} to selected PSUs are given by

$$n_{id}^{AB1} = \left(\frac{\hat{f}_d^{AB1}}{\pi_i^{AB1}} \right) N_{id} \tag{4.2}$$

where $\hat{f}_d^{AB1} = n_{+d}^* \big/ \hat{N}_{+d}^{AB1}$, and $\hat{N}_{+d}^{AB1} = \sum_{i=1}^m N_{id} \big/ \pi_i^{AB1}$. Clearly, $\{n_{id}^{AB1}\}_{1 \le i \le m}$ satisfies n_{+d}^* but does not satisfy n_{i+}^* (or n/m), even in expectation as in $depsem-\mathrm{B1}$, because, in general,

$$E\left(n_{i+}^{AB1}\right) = E\left(\frac{\sum_{d=1}^D \hat{f}_d^{AB1} N_{id}}{\pi_i^{AB1}} \right) = \left(\frac{n}{m}\right) \frac{\sum_{d=1}^D E\left(\hat{f}_d^{AB1}\right) N_{id}}{\sum_{d=1}^D \tilde{f}_d \tilde{N}_{id}} \ne \frac{n}{m}. \tag{4.3}$$

In fact, using Jensen's inequality, it follows that

$$E(n_{i+}^{AB1}) \ge \left(\frac{n}{m}\right) \frac{\sum_{d=1}^D f_d N_{id}}{\sum_{d=1}^D \tilde{f}_d \tilde{N}_{id}} \tag{4.4}$$

where f_d is the domain sampling rate corresponding to the true unknown N_{+d}. Other considerations such as random rounding of n_{id}^{AB1} to obtain integer allocations, the requirement of $n_{id}^{AB1} \le N_{id}$ for all domains within each $i = 1, \ldots, m,$ and the extension to stratified designs (denote by $depsem-\mathrm{AB2}$) carry over in a manner analogous to formula (3.3) for $depsem-\mathrm{B2}$.

***Depsem*–C1 Design:** We propose a $depsem$ design more general than $depsem-\mathrm{AB1}$ for pre-specified π_i $\left(\text{or } \pi_i^{C1}\right)$ when PSU – domain population counts are not known even approximately, so $depsem-\mathrm{AB1}$ is not applicable. As in $depsem-\mathrm{AB1},$ true counts of the PSU – domain sizes N_{id} are obtained through the use of a phase one census of elementary units within selected PSUs. For example, suppose no information about the number of male and female teachers is available for the selected schools. After the schools are selected, we obtain the sex of every teacher in the selected schools for stratification and selection in phase two.

The phase two sample allocations of the desired domain sample sizes to selected PSUs and their properties for $depsem-\mathrm{C1}$ follow easily from those for $depsem-\mathrm{AB1}$. The sample allocations $\{n_{id}^{C1}\}_{1 \le i \le m}$ of the domain total n_{+d}^* for each domain d to selected PSUs are given by

$$n_{id}^{C1} = \left(n_{+d}^*\right) \frac{f_d N_{id} \big/ \pi_i^{C1}}{\sum_{i'=1}^m f_d N_{i'd} \big/ \pi_{i'}^{C1}} = \left(\frac{n_{+d}^*}{\hat{N}_{+d}^{C1}} \right) \frac{N_{id}}{\pi_i^{C1}} = \left(\frac{\hat{f}_d^{C1}}{\pi_i^{C1}} \right) N_{id}, \tag{4.5}$$

where $\hat{N}_{+d}^{C1} = \sum_{i=1}^m N_{i'd} \big/ \pi_{i'}^{C1}$ as the unspecified f_d cancels out. Here, we can set $\hat{f}_d^{C1}$, the domain-level sampling rate, as $n_{+d}^* \big/ \hat{N}_{+d}^{C1}$. As before, an extension to stratified designs (denote by $depsem$ – C2) carries over in a manner analogous to formula (3.3) for $depsem-\mathrm{B2}$.

5 Generalizations of $depsem-\mathrm{C}$ designs in the presence of subsampling within PSUs and nonresponse at the first phase

Often in practice there is subsampling of elementary units within selected PSUs in the first phase because conducting a census of each selected PSU for domain classification may be too costly. In this section, we will consider further generalizations of $depsem-\mathrm{C1}$ where the within $-\mathrm{PSU}$ domain totals are estimated through a sub-sample of elementary units in the first phase rather than determining the $\mathrm{PSU}-$ domain totals exactly (i.e., by census) after the first stage selection. The allocation formulas will differ because we need to take the first phase sampling probabilities of elementary units within selected PSUs into account. Given a selected PSU, let g_i denote the conditional probability of selection for any elementary unit in the first phase sample in PSU $i,$ assuming equal selection probabilities within each PSU. The phase one sample sizes within PSUs are not pre-determined, so the phase one sample should be as large as the schedule and budget allow to maximizing the frames for phase two sampling, especially in PSUs with higher numbers and concentrations of rarer domains.

In addition, up to now the $depsem$ designs have ignored nonresponse at the first phase. If response rates are equal across all PSUs, then the evidence suggests that response probabilities are approximately equal, as well, and can be ignored in specifying suitable sample allocations. However, if response rates vary considerably, then the observed response rate (r_i) in PSU i as an estimated first phase response propensity (assumed to be uniform for all units within PSU i) can be built into the allocation of sample units at the second phase. Building in the response propensity in sample allocation is equivalent to adjusting the phase one selection probabilities for nonresponse, and then a $depsem$ design can be constructed by suitably specifying the domain sampling rates. The design denoted $depsem-\mathrm{C1}'$ includes both sampling of phase one elementary units and phase one nonresponse; this variation was included at the suggestion of Eltinge (2011). Finally, we expand $depsem-\mathrm{C1}'$ to include stratification of PSUs resulting in $depsem-\mathrm{C2}'$ design.

***Depsem* – C1′ Design:** The sample allocations $\{n_{id}^{C1'}\}_{1\le i\le m}$ of the domain total n_{+d}^{*} to selected PSUs, analogous to formula (3.1) of $depsem-\mathrm{B1},$ are given by

$$n_{id}^{C1'} = \left(\frac{\hat{f}_d^{\,C1'}}{\pi_i^{C1'}}\right) N'_{id}, \tag{5.1}$$

where N'_{id} is the size of domain d among phase one sample respondents within PSU $i,$ $\pi_i^{C1'} = \pi_i^{C1} g_i r_i,$ where the unconditional probability of selection for a unit to be in the phase one sample is now $\pi_i^{C1} g_i,$ modified from $depsem-\mathrm{C1}$ due to subsampling in the first phase, $\hat{f}_d^{\,C1'} = n_{+d}^{*} \big/ \hat{N}_{+d}^{C1'},$ and $\hat{N}_{+d}^{C1'} = \sum_{i=1}^{m} N'_{id} \big/ \pi_i^{C1'}.$ Notice that if $r_i = r;$ i.e., equal response rates across all PSUs, then it cancels out in equation (5.1) and has no impact on the sample allocation. Clearly, $\{n_{id}^{C1'}\}_{1\le i\le m}$ satisfies n_{+d}^{*} but does not satisfy n_{i+}^{*} (or n/m) even in expectation as in $depsem-\mathrm{B1}.$

***Depsem* – C2′ Design:** The formula (5.1) can be generalized in a natural way to the stratified case similar to formula (3.1). This case is used in the application considered in the next section. In particular, the sample allocations $\{n_{hid}^{C2'}\}_{1\le i\le m_h, 1\le h\le H}$ of the domain total n_{++d}^{*} to selected PSUs within each stratum h, analogous to formula (3.3) of $depsem-\mathrm{B2}$, are given by

$$n_{hid}^{C2'} = \left(\frac{\hat{f}_d^{C2'}}{\pi_{hi}^{C2'}}\right) N'_{hid}, \tag{5.2}$$

where N'_{hid} is the first phase respondent sample size for domain d within PSU i and stratum $h, \pi_{hi}^{C2'} = \pi_{hi}^{C2} g_{hi} r_{hi}$, where terms are defined in a natural way for stratified designs, $\hat{f}_d^{C2'} = n_{++d}^{*} \big/ \hat{N}_{++d}^{C2'}$, and $\hat{N}_{++d}^{C2'} = \sum_{h=1}^{H} \sum_{i=1}^{m_h} N'_{hid} \big/ \pi_{hi}^{C2'}$.

6 Application of $depsem-\mathrm{C2}'$ design to toolbox development

A team of university researchers developed a set of tests for behavioral and cognitive functions. They desired to “norm” the tests, establishing typical ranges of results for the general population, by measuring the results on children recruited to take the tests. Because the test results vary by age and gender, the goal was to recruit male and female children by year of age. Furthermore, the researchers wanted Spanish-speaking children as well as English-speaking children. The desired domain sample sizes n_{+d}^{*} of completes for twelve age/gender/language cells or domains are shown in Table 6.1.

Table 6.1
Desired completed tests (n_{++d}^{*}) by demographic domain

	English-speaking		Spanish-speaking	
Age	**male**	**female**	**male**	**female**
3	200	200	200	200
4	200	200	200	200
5	200	200	200	200

Originally the researchers desired a probability sample representative of the U.S. population for each of these domains (as well as many additional age groups, which we omit here for simplicity). Once recruited, the sample children were required to be brought to a test site to take the tests in person. Therefore, an area probability design with a limited number of test sites was an efficient design of choice. NORC proposed to select a subsample of the PSU geographies in NORC’s National Frame (Harter, Eckman, English and O’Muircheartaigh 2010). The National Frame is a multi-stage cluster sample of geographies, with housing unit addresses compiled for the smallest level of geography in the sample. The geographies are sampled and the address lists are compiled following the decennial census to support face-to-face interviews throughout the decade.

For norming the tests, 16 of the National Frame’s 79 highest level geographies were selected as PSUs. The population of PSUs was stratified in the same way as the National Frame had been stratified, basically by metropolitan statistical area (MSA) status and size. The strata and sample sizes of PSUs are shown in Table 6.2. For the National Frame, stratum 1 MSAs had been selected with certainty. For the proposed design, the National Frame PSUs within strata were subsampled systematically with pps, where the measure of size was the number of Spanish-speaking households, because the cells for Spanish-speaking children would be the hardest to fill. Probabilities of selection for the PSUs were the product of the original National Frame probabilities and the subsampling probabilities. Some of the Stratum 1 PSUs were subsampled with certainty.

Table 6.2
Subsampling of PSUs from NORC's National Frame

Stratum h	Population #PSUs	National Frame #PSUs	Sample #PSUs (m_h)
1. Largest MSAs	24	24	12
2. Other MSAs	607	17	2
3. Non-MSA Counties	1,852	38	2
Total	2,483	79	16

Each PSU was to be divided into smaller geographical 'site areas'. Each site area would contain a testing site, and the site areas were to be approximately 10×10 miles in urban areas and 30×30 miles in rural areas to provide reasonable driving distances for children to be brought to a test site. Figures 6.1 and 6.2 illustrate the process of defining site areas. In Figure 6.1, a 10×10 mile grid is placed over the Chicago MSA. Then each census tract in the Chicago MSA is assigned to a grid cell based on the geographic location of the tract centroid. The resulting site areas are shown in Figure 6.2. One site area was to be selected per PSU, using systematic pps sampling where the measure of size was the number of Spanish-speaking households. Therefore, in subsequent notation, subscript i denotes both the PSU and the site area.

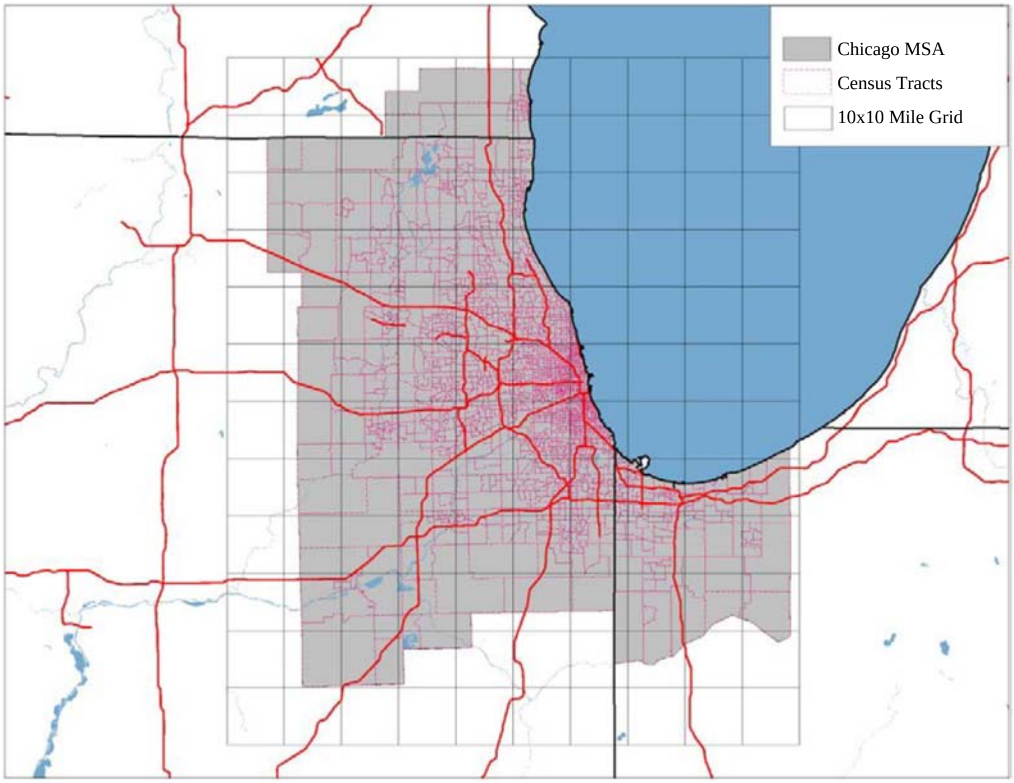


Figure 6.1 10×10 Mile grid over Chicago MSA.

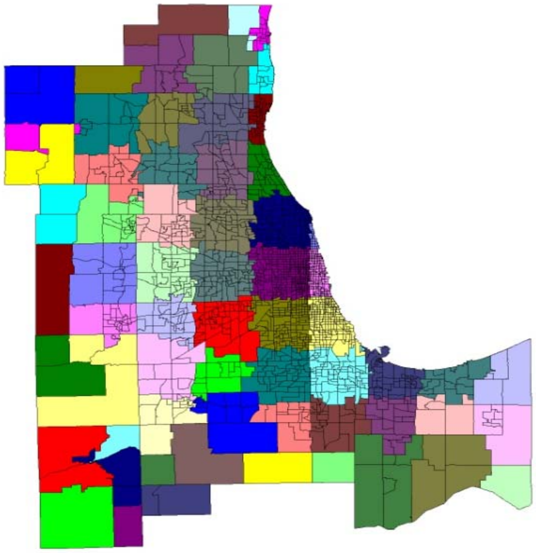


Figure 6.2 Site areas in Chicago MSA with census tracts assigned to grid cells.

We had no information on the number of English-speaking and Spanish-speaking male and female children by year of age for the selected site areas. This problem is best represented by the $depsem - \mathrm{C}2'$ design. Using an address-based sampling frame based on the U.S. Postal Service's Delivery Sequence File, we planned to select a large phase one sample of housing units for a mail screener to roster the households' children by gender, age, and language. The screener also would solicit telephone numbers for contacting parents to gain cooperation for testing the children. In this way we planned to obtain the phase one response rates r_{hi} and the phase two domain frame totals N'_{hid} for each site area i in stratum h.

With the phase one response rates, selection probabilities, and N'_{hid} frame totals in hand, and the specified sample sizes by domain, we were prepared to allocate the desired samples by domain and site area for the second phase of the study to conduct the behavioral and cognitive tests. We would recruit by telephone, with incentives for the sample participants to be brought to the test site.

Ultimately the sample design was never implemented, although we had subsampled the PSUs from the National Frame. Limitations in grant funding led the researchers to revert to convenience sampling near their network of cooperating universities. Nevertheless, the original plan for a probability sample allowed the original Folsom et al. (1987) result for $depsem$ samples to be generalized in a concrete way. For the sake of illustration, we continue the stratified two-phase two stage example with somewhat realistic but hypothetical probabilities and results.

Table 6.3 shows illustrative probabilities of selection for 16 test sites. These hypothetical unconditional site area (PSU) probabilities $\pi_{hi}^{C2'}$ reflect the initial National frame probabilities, the subsampling probabilities for PSUs, and the selection of one test site per PSU. The g_{hi} values are the conditional probabilities of selecting a phase one sample address in stratum h. The product, then, is the unconditional probability of selecting a housing unit (HU) for phase one. Table 6.3 also shows hypothetical site-level response rates r_{hi}, leading to the probabilities $\pi_{hi}^{C2'}$ of an address being selected in phase one and the corresponding household responding and being available for phase two, if eligible.

Suppose that we mailed questionnaires to selected addresses in the site areas to collect household rosters and telephone numbers. Table 6.4 illustrates hypothetical expected counts N'_{hid} (shown as top entries in each cell) by stratum/site area by domain across all 16 test sites. These counts are not true population counts, but they define second phase frame or population counts for our phase two sampling and are based on first phase screener responses.

Table 6.3
Probabilities of phase 1 completion incorporating subsampling and nonresponse

Stratum/PSU (hi)	Unconditional Site Area Probability $\pi_{hi}^{C2} \times 10^6$	Conditional Phase One Sampling rate g_{hi}	Unconditional Phase One Probability $\pi_{hi}^{C2} g_{hi} \times 10^6$	Household Response Rate Per Site r_{hi}	Probability for Phase One Completion $\pi_{hi}^{C2'} \times 10^6$
(1,1)	1,239	0.60	743	0.40	297
(1,2)	972	1.00	972	0.50	486
(1,3)	3,408	0.60	2,045	0.30	613
(1,4)	3,561	0.60	2,137	0.50	1,068
(1,5)	1,985	0.60	1,191	0.40	476
(1,6)	2,083	0.60	1,250	0.40	500
(1,7)	3,142	0.60	1,885	0.60	1,131
(1,8)	5,058	0.60	3,035	0.50	1,517
(1,9)	3,001	0.60	1,801	0.60	1,080
(1,10)	1,621	0.60	973	0.40	389
(1,11)	1,081	0.60	648	0.30	194
(1,12)	533	1.00	533	0.50	266
(2,1)	686	1.00	686	0.40	274
(2,2)	77	1.00	77	0.40	31
(3,1)	328	1.00	328	0.60	197
(3,2)	2,555	0.60	1,533	0.50	766

Table 6.4
Eligible children (N'_{hid}) and sample allocation $\left(n_{hid}^{C2'}\right)$ by stratum, site area, and domain
(Phase two sampling frame counts (top entry) with sample size (bottom entry)
E=English-speaking HU, S=Spanish-speaking HU, M=Male, F=Female, A3=Age 3, A4=Age 4, A5=Age 5)

	Stratum by Site Area (PSU); i.e., (h, i)							
Domain (d)	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)	(1,7)	(1,8)
(E,M,A3)	311 18.8	18 0.7	254 7.4	140 2.4	47 1.8	187 6.7	113 1.8	221 2.6
(E,M,A4)	297 18.4	20 0.8	281 8.4	151 2.6	34 1.3	182 6.7	149 2.4	180 2.2
(E,M,A5)	338 19.0	27 0.9	329 9.0	164 2.6	56 2.0	230 7.7	178 2.6	234 2.6
(E,F,A3)	299 19.3	20 0.8	248 7.8	135 2.4	41 1.7	158 6.1	65 1.1	218 2.8
(E,F,A4)	317 19.8	16 0.6	252 7.6	155 2.7	38 1.5	153 5.7	45 0.7	212 2.6
(E,F,A5)	335 19.7	11 0.4	338 9.7	173 2.8	72 2.6	180 6.3	106 1.6	232 2.7
(S,M, A3)	56 35.0	70 26.8	63 19.1	54 9.4	52 20.3	29 10.8	90 14.8	11 1.3
(S,M,A4)	56 34.9	83 31.6	54 16.3	43 7.4	41 15.9	23 8.5	72 11.8	7 0.9
(S,M,A5)	61 37.9	68 25.8	50 15.0	47 8.1	41 15.9	32 11.8	86 14.0	11 1.3
(S,F,A3)	52 33	74 28.7	61 18.7	41 7.2	29 11.5	23 8.7	95 15.8	14 1.7
(S,F,A4)	63 41.2	79 31.6	56 17.7	52 9.5	29 11.8	25 9.7	79 13.6	11 1.4
(S,F,A5)	54 33.9	86 33.0	61 18.6	50 8.7	16 6.3	32 12.0	81 13.4	11 1.4
Column Margins	2,239 330.9	572 181.7	2,047 155.3	1,205 65.8	496 92.6	1,254 100.7	1,159 93.6	1,362 23.5

Table 6.4 (cont.)
Eligible children $\left(N'_{hid}\right)$ and sample allocation $\left(n^{C2'}_{hid}\right)$ by stratum, site area, and domain
(Phase two sampling frame counts with sample size underneath;
E=English-speaking HU, S=Spanish-speaking HU, M=Male, F=Female, A3=Age 3, A4=Age 4, A5=Age 5)

	Stratum by Site Area (PSU; i.e., (h, i))								**Row**
Domain (d)	**(1,9)**	**(1,10)**	**(1,11)**	**(1,12)**	**(2,1)**	**(2,2)**	**(3,1)**	**(3,2)**	**Margins**
(E,M,A3)	209	252	189	99	52	191	36	7	2,326
	3.5	11.6	17.5	6.7	3.4	111.7	3.3	0.2	200.1
(E,M,A4)	191	221	198	113	63	182	38	5	2,305
	3.3	10.5	18.8	7.8	4.2	109.0	3.6	0.1	200.1
(E,M,A5)	252	234	205	117	65	198	32	9	2,668
	3.9	10.1	17.6	7.4	4.0	107.8	2.7	0.2	200.1
(E,F,A3)	198	236	191	110	63	173	34	5	2,194
	3.5	11.7	18.9	7.9	4.4	108.2	3.3	0.1	200
(E,F,A4)	205	234	187	97	68	185	29	2	2,195
	3.5	11.2	17.9	6.8	4.6	111.9	2.7	0	199.8
(E,F,A5)	245	261	223	124	61	180	41	2	2,584
	4.0	11.8	20.1	8.2	3.9	102.5	3.6	0	199.9
(S,M, A3)	36	45	18	5	11	0	0	27	567
	6.2	21.5	17.2	3.5	7.5	0	0	6.6	200
(S,M,A4)	27	43	25	7	16	0	0	34	531
	4.6	20.5	23.8	4.9	10.8	0	0	8.2	200.1
(S,M,A5)	41	38	27	2	18	0	0	25	547
	7.0	18.0	25.6	1.4	12.1	0	0	6	199.9
(S,F,A3)	34	54	23	7	14	0	0	23	544
	5.9	26.2	22.3	5.0	9.6	0	0	5.7	200
(S,F,A4)	36	50	18	0	11	0	0	25	534
	6.5	25.0	18.0	0	7.8	0	0	6.3	200.1
(S,F,A5)	38	43	29	5	11	0	2	20	539
	6.6	20.6	27.8	3.5	7.5	0	1.9	4.9	200.1
Column Margins	1,512	1,711	1,333	686	453	1,109	212	184	17,534
	58.5	198.7	245.5	63.1	79.8	651.1	21.1	38.3	2,400.2

Using terms defined for equation (5.2), we computed estimated domain counts $\left(\hat{N}^{C2'}_{++d}\right)$ from the first phase sample as shown in Table 6.5. The desired initial domain sample sizes $\left(n^{*}_{++d}\right)$ in Table 6.1 divided by the $\hat{N}^{C2'}_{++d}$ values in Table 6.5 give us the estimated overall sampling rate $\hat{f}^{C2'}_{d}$ for each domain, also shown in Table 6.5.

Again using equation (5.2), we determined the allocations for each stratum, each site area, and each domain within each site area. The resulting allocations are also shown in Table 6.4 as bottom entries in each cell. The allocations are not integers, but random rounding can be used to preserve the *epsem* property in expectation while converting the allocations to integers, as discussed in Section 2. Alternatively, simple rounding will lead to an approximately *depsem* sample design.

Table 6.5
Estimated domain counts $\hat{N}^{C2'}_{++d}$ (in 000)
(Sampling rates ($\hat{f}^{C2'}_{d}$) for phase 2 underneath)

	English-Speaking		**Spanish-speaking**	
Age	**male**	**female**	**male**	**female**
3	11,122	10,399	1,075	1,061
	0.0178	0.0192	0.1860	0.1885
4	10,858	10,749	1,081	1,029
	0.0184	0.0186	0.1851	0.1944
5	11,948	11,415	1,084	1,071
	0.0167	0.0175	0.1845	0.1867

7 Summary and concluding remarks

In the design of any survey, there is a need for good representation of analysis domains in the sample. The sample allocation is not that simple because, unlike information about indicators for commonly used strata available in the sampling frame, domain indicators are generally not available or even if available, it may not be practical to stratify by domains due to interviewer travel costs for in-person surveys. What is needed is a method of sample allocation which allows for desired over-(under-) sampling of domains such that the resulting design is self-weighting or *epsem* for domains. Such designs are desirable for variance efficiency in general. In the case of one phase two stage designs, under certain assumptions, it is possible to allocate equal interviewer workload per selected PSU such that the sample size for all selected PSUs is controlled at the desired level, but domain sizes over all selected PSUs satisfy the desired level only in expectation. On the other hand, in the case of two phase two stage designs, it is possible to allocate domain sample sizes within PSUs such that the domain sizes over all PSUs are controlled at desired levels but the sample size per selected PSU is not controlled as the equal interviewer workload per PSU is not deemed important in this case. Although the *epsem* design of Kish at the population level is well known, domain level *epsem* (denoted *depsem* in this paper) designs are not well known among practitioners.

In this paper, we considered two main scenarios for *depsem* designs considered by Folsom et al. (1987). First, for two stage designs with known domain level PSU population counts (as well as known frame-level domain identifiers for elementary units) and pre-specified domain sample sizes, the PSU selection probabilities are defined such that the desired PSU sample size (equal per PSU) is allocated to domains within PSUs to obtain a *depsem* design. Second, for two phase two stage designs with known PSU selection probabilities and pre-specified domain sample sizes, the domain sampling rates are defined such that the desired domain sample size is allocated to PSUs within domains to obtain a *depsem* design. These two designs were referred to as $depsem-\mathrm{A1}$ and $\mathrm{B1}$ respectively. A simple justification of these two designs was provided. It is based on the key idea for obtaining *depsem* designs that the sampling rate (n_{id}/N_{id}) at the PSU by domain level should be made directly proportional to the domain level sampling rate (f_d) but inversely proportional to the PSU selection probability (π_i). For $depsem-\mathrm{A1}$, f_d is known but π_i is suitably defined (it was termed *composite measure of size* by Folsom et al. 1987), while for $depsem-\mathrm{B1}$, π_i is known but f_d is suitably defined. The corresponding stratified versions (denoted by $\mathrm{A2}/\mathrm{B2}$) can also be easily defined.

As a generalization of $depsem-\mathrm{A1}/\mathrm{B1}$ designs, $depsem-\mathrm{AB1}$ was proposed where domain-level PSU population counts are only approximately known for specifying PSU selection probabilities, but a two phase design is used to allocate desired domain sample sizes to PSUs after obtaining the true domain-level population counts for selected PSUs in the first phase. Also generalizations of $depsem-\mathrm{B1}$ were considered to obtain $depsem-\mathrm{C1}$ when PSU selection probabilities are pre-specified from other considerations. The $depsem-\mathrm{C1}'$ extends $depsem-\mathrm{C1}$ to cover certain practical realistic situations: 1) subsampling of elementary units within each selected PSU in the first phase to reduce cost, and 2) nonresponse in screening units for domain classification. The $depsem-\mathrm{C2}'.$ design allows for stratification in addition to practical features of $depsem-\mathrm{C1}'$ mentioned above. For all *depsem* designs except for $\mathrm{A1},$ PSU sample size is not directly controlled, but domain sample size is controlled via stratification of the first phase before the second phase. This is not a limitation in various practical applications where interviews are not conducted face-to-face.

The initial *depsem* design framework of Folsom et al. (1987) to allocate equal probability samples for multiple domains in two-stage designs in conjunction with one/two phase is a useful technique currently available in the SUDAAN software system (http://www.rti.org/page.cfm/SUDAAN) and employed successfully at RTI International for many years for studies such as the National Survey of Child and Adolescent Well-Being. The generalizations presented here extend the technique to the situation of multiple domains where the domain-level population counts need to be estimated for all selected PSUs, and where PSU selection probabilities are pre-specified from other considerations. These techniques are expected to be useful to sampling statisticians in a variety of situations.

Acknowledgements

The first author is grateful to Ralph Folsom of RTI International for several useful discussions. The authors thank Ned English of NORC at the University of Chicago for his help in creating site areas for the toolbox application. Thanks are also due to the associate editor and two referees for their very useful comments which helped improve clarity and presentation of the paper.

References

Cochran, W. (1977). *Sampling Techniques,* 3rd Ed. New York: John Wiley & Sons, Inc.

Eltinge, J. (2011). Personal Communication.

Fahimi, M., and Judkins, D. (1991). PSU Probabilities given differential sampling at second stage. *Proceedings of the Section on Survey Research Methods,* American Statistical Association, 538-543.

Folsom, R.E., Potter, F.J. and Williams, S.R. (1987). Notes on a composite size measure for self-weighting samples in multiple domains. *Proceedings of the Section on Survey Research Methods,* American Statistical Association, 792-796.

Harter, R., Eckman, S., English, N. and O'Muircheartaigh, C. (2010). Applied sampling for large-scale multi-stage area probability designs. In *Handbook of Survey Research, Second Edition,* (Eds., P. Marsden and J. Wright), Emerald: United Kingdom.

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

Lohr, S. (2010). *Sampling: Design and Analysis,* 2nd Ed. Boston: Brooks/Cole.

Singh, A.C., and Harter, R.M. (2011). A generalized epsem two-phase design for domain estimation. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 3269-3282.

Survey Methodology, December 2015
Vol. 41, No. 2, pp. 315-331
Statistics Canada, Catalogue No. 12-001-X

A design effect measure for calibration weighting in single-stage samples

Kimberly A. Henry and Richard Valliant[1]

Abstract

We propose a model-assisted extension of weighting design-effect measures. We develop a summary-level statistic for different variables of interest, in single-stage sampling and under calibration weight adjustments. Our proposed design effect measure captures the joint effects of a non-*epsem* sampling design, unequal weights produced using calibration adjustments, and the strength of the association between an analysis variable and the auxiliaries used in calibration. We compare our proposed measure to existing design effect measures in simulations using variables like those collected in establishment surveys and telephone surveys of households.

Key Words: Auxiliary data; Kish weighting design effect; Spencer design effect; Generalized regression estimator.

1 Introduction

A design effect (deff) in its general form measures the relative increase or decrease in the variance of an estimator due to departures from simple random sampling. Kish (1965) presented the deff as a convenient way of gauging the effect of clustering on an estimator of a mean. Park and Lee (2004) review some of the history behind the formulation and use of deff's. Design effects are especially useful in approximating the total sample size needed in a cluster sample. Clustering usually causes some loss of efficiency and the variance from a simple random sample, which is easy to compute, can be multiplied by a deff to approximate the variance that would be obtained from a cluster sample. This can, in turn, be used to determine the total sample size needed in a cluster sample to achieve a desired level of precision. Later work by Rao and Scott (1984) and others found that more complicated versions of deff's were useful to adjust inferential statistics calculated from complex survey data.

A specialized version of the deff was proposed in Kish (1965) that addressed only the effect of using weights that are not all equal. Kish derived the "design effect due to weighting" for a case in which weights vary for reasons other than statistical efficiency. On the other hand, there are sample designs and estimators where having varying weights can be quite efficient. An establishment survey where population variances of analysis variables differ markedly among industries is one example. Calibrating to population counts can also produce different sized weights but is an essential tool in attempting to correct for coverage errors in some surveys, like ones done by telephone. Spencer (2000) proposed a simple model-assisted approach to estimate the impact on variance of using variable weights in a situation where an analysis variable depends on a single covariate.

The Kish and Spencer measures, reviewed in Section 2, do not provide a summary measure of the impact of the gains in precision that may accrue from sampling with varying probabilities and using a calibration estimator like the general regression (GREG) estimator. While the Kish design effects attempt to measure the impact of variable weights, they are informative only under special circumstances, do not

[1] Kimberly A. Henry, U.S. Internal Revenue Service, Washington DC. E-mail: kimberly.a.henry@irs.gov; Richard Valliant, Universities of Michigan & Maryland, College Park MD 20854. E-mail: rvalliant@survey.umd.edu.

account for alternative variables of interest, and can incorrectly measure the impact of differential weighting in some circumstances, facts noted in Kish (1992). Survey practitioners should be cautious when using this measure in informative sampling and estimation schemes in which there exists an intentional relationship between the weights and variables of interest. Spencer's approach holds for with-replacement single-stage sampling for a very simple estimator of the total constructed with inverse-probability weights with no further adjustments. There are also few empirical examples comparing these measures in the literature.

Calibration adjustments are often applied to reduce variances and correct for undercoverage and/or nonresponse in surveys (e.g., Särndal and Lundström 2005; Kott 2009). When the calibration covariates are correlated with the coverage/response mechanism, calibration weights can improve the mean squared error (MSE) of an estimator. In many applications, since calibration involves unit-level adjustments, calibration weights can vary more than the base weights or category-based nonresponse or poststratification adjustments (Kalton and Flores-Cervantes 2003; Brick and Montaquila 2009). Thus, an ideal measure of the impact of calibration weights incorporates not only the correlation between the survey variable of interest y and the weights, but also the correlation between y and the calibration covariates $\mathbf{x}$ to avoid "penalizing" weights for the mere sake that they vary.

In Section 3, we introduce a new design effect measure that accounts for the joint effect of a non-epsem sample design and unequal weight adjustments in the larger class of calibration estimators. It is assumed that a probability sample design is used and that there are no missing data problems that would induce a dependence between sample inclusion and the values of the y's. Our summary measure incorporates the survey variable, using a generalized regression variance to reflect multiple calibration covariates. In Section 4, we apply the estimators in a simulation using variables similar to ones collected in establishment surveys and household surveys done by telephone and demonstrate empirically how the proposed estimator outperforms the existing methods in the presence of unequal calibration weights. Section 5 is a conclusion.

2 Existing methods

In this section, we specify notation and summarize the Kish and Spencer measures. The assumptions used to derive each of these are also presented.

2.1 GREG weight adjustments

Case weights resulting from calibration on benchmark auxiliary variables can be defined with a global regression model for the survey variables (see Kott 2009 for a review). Deville and Särndal (1992) proposed the calibration approach that involves minimizing a distance function between the base weights and final weights to obtain an optimal set of survey weights. Specifying alternative calibration distance functions produces alternative estimators. Suppose that a single-stage probability sample of n units is selected with π_i being the selection probability of unit i and $\mathbf{x}_i$ a vector of p auxiliaries associated with unit $i.$ A least squares distance function produces the *general regression estimator* (GREG):

$$\hat{T}_{\text{GREG}} = \hat{T}_{\text{HT}y} + \hat{\mathbf{B}}^T \left(\mathbf{T}_x - \hat{\mathbf{T}}_{\text{HT}x}\right) = \sum\nolimits_{i \in s} g_i y_i \big/ \pi_i \,, \tag{2.1}$$

where $\hat{T}_{\mathrm{HT}y} = \sum_{i \in s} y_i / \pi_i$ is the Horvitz-Thompson (HT 1952) estimator of the population total of $y, \hat{\mathbf{T}}_{\mathrm{HT}x} = \sum_{i \in s} \mathbf{x}_i / \pi_i$ is the vector of HT estimated totals for the auxiliary variables, $\mathbf{T}_x = \sum_{i=1}^{N} \mathbf{x}_i$ is the corresponding vector of known totals, $\hat{\mathbf{B}} = \mathbf{A}_s^{-1} \mathbf{X}_s^T \mathbf{V}_{ss}^{-1} \mathbf{\Pi}_s^{-1} \mathbf{y}_s$ is the regression coefficient, with $\mathbf{A}_s = \mathbf{X}_s^T \mathbf{V}_{ss}^{-1} \mathbf{\Pi}_s^{-1} \mathbf{X}_s$, $\mathbf{X}_s^T$ is the matrix of $\mathbf{x}_i$ values in the sample, $\mathbf{V}_{ss} = \mathrm{diag}(v_i)$ is the diagonal of the variance matrix specified under the working model $y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \varepsilon_i \sim (0, v_i)$, and $\mathbf{\Pi}_s = \mathrm{diag}(\pi_i)$. In the second expression for the GREG estimator in (2.1), $g_i = 1 + (\mathbf{T}_x - \hat{\mathbf{T}}_{\mathrm{HT}x})^T \mathbf{A}_s^{-1} \mathbf{x}_i v_i^{-1}$ is the "$g-$weight," such that the case weights are $w_i = g_i / \pi_i$ for each sample unit i.

The GREG estimator for a total is model-unbiased under the associated working model. The GREG is consistent and approximately design-unbiased when the sample size is large (Särndal, Swensson and Wretman 1992). When the model is correct, the GREG estimator achieves efficiency gains. If the model is incorrect, then the efficiency gains will be dampened (or nonexistent) but the GREG estimator is still approximately design-unbiased. Relevant to this work, the variance of the GREG estimator can be used to approximate the variance of any calibration estimator (Deville and Särndal 1992; Deville, Särndal and Sautory 1993) when the sample size is large. This allows us to produce one design effect measure applicable to all estimators in the family of calibration estimators.

2.2 The direct design-effect measures for single-stage samples

For a given $\mathrm{non-epsem}$ sample π and estimator $\hat{T}$ for the finite population total T, one definition for the *direct design effect* (Kish 1965) is

$$\mathrm{Deff}(\hat{T}) = \mathrm{Var}_\pi(\hat{T}) / \mathrm{Var}_{\mathrm{srswr}}(\hat{T}_{\mathrm{srswr}}) \tag{2.2}$$

where $\hat{T}_{\mathrm{srswr}}$ is the estimator of a total based on a simple random sample selected with replacement (srswr). We refer to this as a "direct" population quantity since it uses theoretical variances in the numerator and denominator. The design effect in (2.2) measures the size of the variance of the estimator $\hat{T}$ under the design π, relative to the variance of the estimator of the same total if a srswr of the same size had been used.

In large samples, we can approximate the variance of any calibration estimator $\hat{T}_{\mathrm{cal}}$ using the approximate variance of the GREG (GREG AV, Särndal et al. 1992; Deville et al. 1993), such that the design effect is

$$\mathrm{Deff}(\hat{T}_{\mathrm{cal}}) \doteq \mathrm{Var}_\pi(\hat{T}_{\mathrm{GREG}}) / \mathrm{Var}_{\mathrm{srswr}}(\hat{T}_{\mathrm{srswr}}). \tag{2.3}$$

To estimate these design-effects, we use the appropriate corresponding sample-based variance estimates. Estimates of both measures (2.2) and (2.3) can be produced using conventional survey estimation software. Our proposed design effect is a model-assisted approximation to (2.3).

2.3 Kish's "Haphazard-sampling" design-effect measure for unequal weights

Kish (1965, 1990) proposed the "design effect due to weighting" as a measure to quantify the loss of precision due to using unequal and inefficient weights. For $\mathbf{w} = (w_1, \ldots, w_n)^T$, this measure is

$$\begin{aligned}\operatorname{deff}_K(\mathbf{w}) &= 1+[\mathrm{CV}(\mathbf{w})]^2\\ &= \frac{n\sum_{i\in s} w_i^2}{\left[\sum_{i\in s} w_i\right]^2},\end{aligned} \tag{2.4}$$

where $\mathrm{CV}(\mathbf{w}) = \sqrt{n^{-1}\sum_{i\in s}(w_i-\bar{w})^2 / \bar{w}^2}$ is the coefficient of variation of the weights with $\bar{w} = n^{-1}\sum_{i\in s} w_i$. Expression (2.4) is derived from the ratio of the variance of the weighted survey mean under disproportionate stratified sampling to the variance under proportionate stratified sampling when all stratum unit variances are equal (Kish 1992). With equal stratum variances, sampling with a proportional allocation to strata is optimal, which leads to all units having the same weight.

Kish referred to (2.4) as a measure that is appropriate for “haphazard” weighting in which unequal weights are inefficient. Kish (1992) and Park and Lee (2004) give examples of informative sampling where this measure does not apply. Park and Lee (2004) also demonstrate this measure may not apply equally well to estimators of means and totals.

2.4 Spencer’s model-assisted measure for PPSWR sampling

Spencer (2000) derives a design-effect measure to more fully account for the effect on variances of weights that are correlated with the survey variable of interest. The sample is assumed to be selected with varying probabilities and with replacement (denoted as PPSWR sampling here). A particular case of this would be $p_i \propto x_i$, where x_i is a measure of size associated with unit i and p_i is the one-draw probability of selecting unit i. Suppose that p_i is correlated with y_i and that a linear model holds for $y_i : y_i = \alpha + \beta p_i + \varepsilon_i$. If the entire finite population were available, then the ordinary least squares estimates of α and β are $A = \bar{Y} - B\bar{P}$ and $B = \sum_{i\in U}(y_i - \bar{Y})(p_i - \bar{P}) \big/ \sum_{i\in U}(p_i - \bar{P})^2$, where $\bar{Y}, \bar{P}$ are the finite population means for y_i and p_i. The finite population variance of the residuals, $e_i = y_i - (A + Bp_i)$, is $\sigma_e^2 = (1-\rho_{yp}^2)N^{-1}\sum_{i\in U}(y_i - \bar{Y})^2 = (1-\rho_{yp}^2)\sigma_y^2$, where σ_y^2 is the finite population variance of y and ρ_{yp} is the finite population correlation between y_i and p_i. The estimated total studied by Spencer is referred to as the $\mathrm{pwr}-$estimator or Hansen-Hurwitz (1943) estimator (Särndal et al. 1992, Section 2.9) and is defined as $\hat{T}_{\mathrm{pwr}} = n^{-1}\sum_{i=1}^{n} y_i / p_i$, with design-variance $\mathrm{Var}(\hat{T}_{\mathrm{pwr}}) = n^{-1}\sum_{i\in U} p_i (y_i / p_i - T)^2$ in single-stage sampling. For use below, define $w_i = (np_i)^{-1}$. Spencer substituted the model-based values for y_i into the $\mathrm{pwr}-$estimator’s variance and took its ratio to the variance of the estimated total using srswr to produce the following design effect for unequal weighting (see Appendix in Spencer 2000):

$$\mathrm{Deff}_S = \frac{A^2}{\sigma_y^2}\left(\frac{n\bar{W}}{N} - 1\right) + \frac{n\bar{W}}{N}\left(1-\rho_{yp}^2\right) + \frac{n\rho_{e^2w}\sigma_{e^2}\sigma_w}{N\sigma_y^2} + \frac{2An\rho_{ew}\sigma_e\sigma_w}{N\sigma_y^2} \tag{2.5}$$

where $\bar{W} = N^{-1}\sum_{i\in U} w_i = (nN)^{-1}\sum_{i\in U} 1/p_i$ is the average weight in the population, ρ_{e^2w} and ρ_{ew} are the finite population correlation of the e_i^2’s with the w_i’s and the e_i’s with the w_i’s, respectively; $\sigma_{e^2}^2$

and σ_w^2 are the finite population variances of the e_i^2's and w_i's. In skewed populations, the correlation ρ_{ew} in (2.5) may be negligible but ρ_{e^2w} can be large and negative if units with larger $\mathbf{x}, y$ values have larger residuals but small weights. We found empirically in the simulations reported in Section 4 that ρ_{e^2w} was generally negative and larger in relative size than ρ_{ew}.

Assuming that the correlations in the last two terms of (2.5) are negligible, Spencer approximates (2.5) with

$$\text{Deff}_S \approx \left(1-\rho_{yp}^2\right)\frac{n\bar{W}}{N}+\left(\frac{A}{\sigma_y}\right)^2\left(\frac{n\bar{W}}{N}-1\right), \tag{2.6}$$

A similar expression is given by Park and Lee (2004; expression 4.7). Spencer proposed estimating measure (2.6) with

$$\text{deff}_S = \left(1-R_{yp}^2\right)\text{deff}_K(\mathbf{w})+\left(\hat{\alpha}/\hat{\sigma}_y\right)^2\left(\text{deff}_K(\mathbf{w})-1\right), \tag{2.7}$$

where R_{yp}^2 and $\hat{\alpha}$ are the $R-$squared and estimated intercept from fitting the model $y_i = \alpha+\beta p_i+\varepsilon_i$ with survey weighted least squares, $\hat{\sigma}_y^2 = \sum_{i\in s} w_i\left(y_i-\hat{\bar{y}}_w\right)^2 / \sum_{i\in s} w_i$ with $\hat{\bar{y}}_w = \sum_s w_i y_i / \sum_s w_i$ is the estimated population unit variance. Spencer's estimator (2.7) assumes that the population size N is large.

When ρ_{yp} is zero and σ_y is large, measure (2.7) is approximately equivalent to Kish's measure (2.4). However, Spencer's method does incorporate the survey variable y_i, unlike (2.4), and implicitly reflects the dependence of y_i on the selection probabilities p_i. We can explicitly see this by noting that when N is large, $A=\bar{Y}-BN^{-1}\approx\bar{Y}$, and (2.6) can be written as

$$\text{Deff}_S \approx \left(1-\rho_{yp}^2\right)\frac{n\bar{W}}{N}+\frac{1}{\text{CV}_{\bar{Y}}^2}\left(\frac{n\bar{W}}{N}-1\right), \tag{2.8}$$

where $\text{CV}_{\bar{Y}}^2 = \sigma_y^2/\bar{Y}^2$ is the population-level unit coefficient of variation (CV). We estimate (2.8) with

$$\text{deff}_S = \left(1-R_{yp}^2\right)\text{deff}_K(\mathbf{w})+\frac{1}{\text{cv}_y^2}\left(\text{deff}_K(\mathbf{w})-1\right), \tag{2.9}$$

where $\text{cv}_y^2 = \hat{\sigma}_y^2/\hat{\bar{y}}_w^2$. Note that cv_y is not the standard CV produced in conventional survey estimation software, since it estimates the population unit CV of y.

3 Proposed design-effect measure

We extend Spencer's (2000) approach in single-stage sampling to produce a new weighting design effect for a calibration estimator. While Spencer's assumed $y_i=\alpha+\beta p_i+\varepsilon_i$, we model y_i as $y_i = \alpha+\mathbf{x}_i^T\boldsymbol{\beta}+\varepsilon_i = \dot{\mathbf{x}}_i^T\dot{\boldsymbol{\beta}}+\varepsilon_i$, where $\dot{\mathbf{x}}_i = \left[1 \quad \mathbf{x}_i\right]$ and $\dot{\boldsymbol{\beta}} = [\alpha \quad \boldsymbol{\beta}]$. Denote the full finite population estimators of α and $\boldsymbol{\beta}$ by $A=\bar{Y}-\bar{\mathbf{X}}\mathbf{B}$ and $\mathbf{B}=\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$ where $\mathbf{X}$ is the $N\times p$ matrix of auxiliaries for the N units in the finite population and $\mathbf{Y}$ is the $N-$vector of y values. The finite population residuals are defined as $e_i = y_i-\left(A+\mathbf{x}_i^T\mathbf{B}\right)\equiv y_i-\dot{\mathbf{x}}_i^T\dot{\mathbf{B}}$ where $\dot{\mathbf{B}} = [A \quad \mathbf{B}]$.

Producing the design effect proposed below involves four steps: (1) constructing a linear approximation to the GREG estimator; (2) obtaining the design-variance of this linear approximation; (3) substituting model-based components into the GREG variance; and (4) taking the ratio of this model-assisted variance to the variance of the $\text{pwr}-$estimator of the total under srswr. Since steps $(1)-(4)$ produce the theoretical design effect for an estimator, we add the final step: (5) plug-in sample-based estimates for each theoretical design effect component.

Step 1. A linearization of the GREG estimator (Expression 6.6.9 in Särndal et al. 1992) is

$$\begin{aligned} \hat{T}_{\text{GREG}} &\doteq \hat{T}_{\text{HT}y} + \left(\mathbf{T}_x - \hat{\mathbf{T}}_{\text{HT}x}\right)^T \dot{\mathbf{B}} \\ &= \mathbf{T}_x^T \dot{\mathbf{B}} + \sum\nolimits_{i\in s} e_i \big/ \pi_i \end{aligned} \tag{3.1}$$

where $\sum\nolimits_{i\in s} e_i \big/ \pi_i$ is the HT estimator of the population total of the e_i, $E_U = \sum\nolimits_{i\in U} e_i$. To obtain a simple variance formula in step 2, we treat the case of with-replacement sampling and replace $\sum\nolimits_{i\in s} e_i \big/ \pi_i$ with the $\text{pwr}-$estimator $n^{-1}\sum\nolimits_{i=1}^{n} e_i \big/ p_i$. Next, define δ_i to be the number of times that unit i is selected for the sample. Since $E_\pi(\delta_i) = np_i$, the second component in (3.1) has design-expectation $E_\pi\left(n^{-1}\sum\nolimits_{i=1}^{n} e_i \big/ p_i\right) = E_U$.

Step 2. From step 1 with the assumption of with-replacement sampling, $\hat{T}_{\text{GREG}} - \mathbf{T}_x^T\dot{\mathbf{B}} \doteq n^{-1}\sum\nolimits_{i=1}^{n} e_i \big/ p_i$, with design-variance

$$\begin{aligned} \text{Var}_\pi\left(\hat{T}_{\text{GREG}} - \mathbf{T}_x^T \dot{\mathbf{B}}_U\right) &\doteq \text{Var}_\pi\left(n^{-1}\sum\nolimits_{i=1}^{n} e_i \big/ p_i\right) \\ &= n^{-1}\sum\nolimits_{i\in U} p_i \left(e_i \big/ p_i - E_U\right)^2. \end{aligned} \tag{3.2}$$

Steps 3 and 4. We follow Spencer's approach and substitute model values in variance (3.2) to formulate a design-effect measure. However, we substitute in the model-based equivalent to e_i, not y_i. Substituting the GREG residuals e_i into the variance and taking its ratio to the variance of the $\text{pwr}-$estimator in simple random sampling with replacement, $\text{Var}_{\text{srswr}}\left(\hat{T}_{\text{srswr}}\right) = N^2\sigma_y^2 \big/ n$, where $\sigma_y^2 = N^{-1}\sum\nolimits_{i=1}^{N}\left(y_i - \bar{Y}\right)^2$, will produce our approximate design effect due to unequal calibration weighting. We can simplify things greatly by defining $u_i = A + e_i$, where $u_i = y_i - \mathbf{x}_i^T\mathbf{B}$, which implies $\bar{U} = A + \bar{E}_U = A$. The resulting design effect (see Appendix) is

$$\text{Deff}_H = \frac{n\bar{W}}{N}\left(\frac{\sigma_u^2}{\sigma_y^2}\right) + \frac{n\sigma_w}{N\sigma_y^2}\left(\rho_{u^2w}\sigma_{u^2} - 2A\rho_{uw}\sigma_u\right) \tag{3.3}$$

where $\sigma_u^2 = N^{-1}\sum\nolimits_{i=1}^{N}\left(u_i - \bar{U}\right)^2, \sigma_y^2 = N^{-1}\sum\nolimits_{i=1}^{N}\left(y_i - \bar{Y}\right)^2, \rho_{u^2w}$ is the finite population correlation between u_i^2 and $w_i, \sigma_{u^2}^2$ is the variance of u_i^2 and ρ_{uw} is the correlation between u_i and w_i.

The first component in (3.3) is $O(1)$; the factor $n\bar{W}/N$ is related to the Kish deff as described below. The factor σ_u^2/σ_y^2 is an adjustment based on the effectiveness of the covariates in predicting y. The second component in (3.3) is $O(n/N)$ and incorporates terms related to the strength of the relationship between the calibration covariates and the weights.

Note that the derivation of (3.3) assumes with-replacement (WR) sampling was used. Although without replacement (WOR) sampling is more common in practice, the WOR variance of an estimated total is complicated since it involves joint selection probabilities. The WR variance formula is simple enough to provide insights into the effect of calibration on a deff. In cases where there are gains in precision from using WOR sampling, an ad hoc finite population correction factor can be incorporated in (3.3), i.e., $(1-n/N)\,\mathrm{Deff}_H$.

Step 5. To estimate (3.3), we use

$$\mathrm{deff}_H \approx \mathrm{deff}_K(\mathbf{w})\frac{\hat{\sigma}_u^2}{\hat{\sigma}_y^2} + \frac{n\hat{\sigma}_w}{N\hat{\sigma}_y^2}\left(\hat{\rho}_{u^2w}\hat{\sigma}_{u^2} - 2\hat{\alpha}\hat{\rho}_{uw}\hat{\sigma}_u\right), \tag{3.4}$$

where the model parameter estimate $\hat{\alpha}$ is obtained using survey-weighted least squares, $\hat{\sigma}_y^2$ was defined in Section 2.3, $\hat{\sigma}_u^2 = \sum_{i\in s} w_i\,(\hat{u}_i - \bar{u}_w)^2 / \sum_{i\in s} w_i$, $\hat{u}_w = \sum_{i\in s} w_i\hat{u}_i / \sum_{i\in s} w_i$, $\hat{u}_i = y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}$, and $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}_s^T\mathbf{W}\mathbf{X}_s\right)^{-1}\mathbf{X}_s^T\mathbf{W}\mathbf{y}_s$ is the survey-weighted least-squares estimate of $\boldsymbol{\beta}$, with $\mathbf{W} = \mathrm{diag}(w_1,\ldots,w_n)$, and other terms defined in Section 2.1.

If the correlations in (3.3) are negligible or the sampling fraction n/N is small, the first term dominates and we obtain

$$\mathrm{Deff}_H \approx \frac{n\bar{W}}{N}\left(\frac{\sigma_u^2}{\sigma_y^2}\right),$$

which can be estimated with

$$\mathrm{deff}_H \approx \mathrm{deff}_K(\mathbf{w})\,\hat{\sigma}_u^2/\hat{\sigma}_y^2. \tag{3.5}$$

Note that in samples without calibration weight adjustments, we have $\hat{u}_i = y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}} \approx y_i$ and $\sigma_u^2 \approx \sigma_y^2$. In this case expression (3.5) becomes $\mathrm{Deff}_H \approx n\bar{W}/N$, which we estimate with Kish's measure $\mathrm{deff}_K = 1 + [\mathrm{CV}(\mathbf{w})]^2$. However, when the relationship between the calibration covariates $\mathbf{x}$ and y is stronger, the variance σ_u^2 should be smaller than σ_y^2. In this case, measure (3.5) is smaller than Kish's estimate. Variable weights produced from calibration adjustments are thus not as "penalized" (shown by overly high design effects) as they would be using the Kish and Spencer measures. However, if we have "ineffective" calibration, or a weak relationship between $\mathbf{x}$ and y, then σ_u^2 can be greater than σ_y^2, producing a design effect greater than one. The Spencer measure only accounts for an indirect relationship between $\mathbf{x}$ and y if there was only one x and it was used to produce p_i. This is illustrated in Section 4. We also examine the extent to which the correlation components in our proposed design effect (3.3) are large enough to influence the exact measure. Calculation of (3.3) requires only the sample $y-$values, covariates, and calibration weights. This measure can, thus, be produced more quickly than measure (2.3), whose components are often available later in data processing after a variance estimation system is set up.

4 Empirical evaluation

We conducted two simulation studies using data that mimic single-stage sampling. The first utilizes publically-available data from tax returns and continuous variables of interest, while the second examines the performance of the alternative measures for a binary outcome measure in a single-stage survey.

4.1 Establishment data simulation study

Here a sample dataset of tax return data is used to mimic an establishment survey setup. The data come from the Tax Year 2007 Statistics of Income (SOI) Form 990 Exempt Organization (EO) sample. This is a stratified Bernoulli sample of 22,430 EO tax returns selected from 428,719 filed with and processed by the IRS between December 2007 and November 2009. This sample dataset, along with the population frame data, is free and electronically available online (Statistics of Income 2011). These data make a candidate "establishment-type" dataset for estimating design effects, in which Kish's design effect may not apply.

The SOI EO sample dataset is used here as a pseudopopulation for illustration. Four variables of interest are used: Total Assets, Total Liabilities, Total Revenue, and Total Expenses. Returns that were sampled with certainty or that had "very small" assets (defined by having Total Assets less than $1,000,000, including zero) were removed, leaving 8,914 units. We then randomly replicated and perturbed the data to create a pseudopopulation of 50,000 units. We used simple random sampling with replacement to select more observations, then the additional data values were perturbed using the `jitter` (Chambers, Cleveland, Kleiner and Tukey 1983) function in R.

Figure 4.1 shows a pairwise plot of the pseudo-population, including plots of the variable values against each other in the lower left panels, histograms on the diagonal panels, and the correlations among the variables in the upper right panels. This plot mimics establishment-type data patterns. From the diagonal panels, we see that the variables of interest are all highly skewed. From the lower left panels, there exists a range of different relationships among them. The Total Assets variable is less related to Total Revenue and Total Expenses (with moderate correlations of $0.41-0.44$); Total Revenue and Total Expenses are highly correlated.

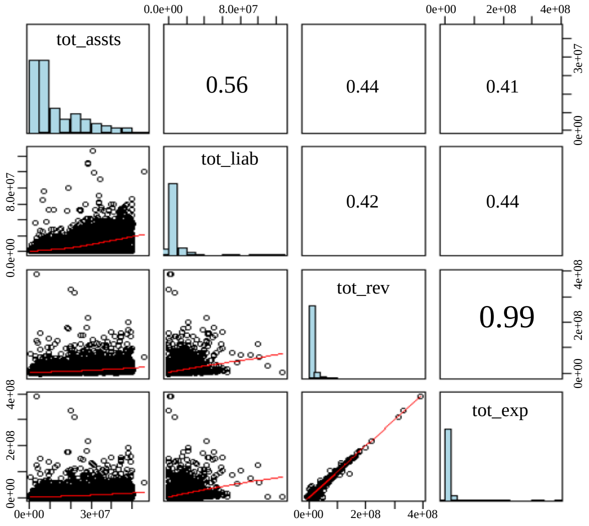


Figure 4.1 Pseudopopulation values and loess lines for design effect evaluation.

Three sizes of samples were selected $(n = 100; 500; 1{,}000)$ without replacement from the pseudopopulation using the square root of Total Assets as a measure of size. This type of sampling is referred to as $\pi\mathrm{ps}$ sampling subsequently. The HT weights were then calibrated using the “`linear`” method in the `calibrate` function in the `survey` package for R (corresponding to a GREG estimator, Lumley 2012) to match the totals of an intercept, Total Assets and Total Revenue. The analysis variables are Total Liabilities and Total Expenses. (Note that we follow the common practice of developing procedures in the previous sections using formulas for with-replacement sampling but empirically evaluating them in without-replacement samples, which are the type used in applications.)

Eight design effects estimates are considered:

- Estimates of the design effect measures (2.2) and (2.3). Expression (2.2) reflects the efficiency of $\pi\mathrm{ps}$ sampling and use of the $\mathrm{HT}-$estimator. Expression (2.3) reflects gains (if any) of $\pi\mathrm{ps}$ sampling combined with GREG estimation;
- The Kish measure (2.4) computed using the GREG weights;
- Three Spencer measures computed using the GREG weights: (i) the exact measure that estimates (2.5), (ii) the approximation (2.7) assuming zero correlation terms, and (iii) the large-population approximation (2.9). The Spencer measures are designed to reflect gains due to PPSWR sampling and use of the $\mathrm{pwr}-$estimator. It does not account for any gains due to calibration.
- Two proposed measures: (i) the exact proposed single-stage design effect (3.4) and (ii) the zero-correlation approximation (3.5). Both of these are meant to show the precision gains (if any) of PPSWR sampling combined with GREG estimation.

Note that neither the Spencer nor the proposed measures account for any reduction in variances due to sampling a large fraction of the population.

We selected ten thousand samples to further understand the empirical behavior of the alternative design effect estimators. The empirical relbiases and ratio of the mean square errors (MSE’s) of the totals are

$$\begin{aligned}\mathrm{relbias}\left(\hat{T}\right) &= 100 \times \sum\nolimits_{s=1}^{S}\left(\hat{T}_s - T\right)\Big/T\\ \text{MSE ratio} &= \mathrm{MSE}\left(\hat{T}_{\mathrm{HT}}\right)\Big/\mathrm{MSE}\left(\hat{T}_{\mathrm{GREG}}\right)\\ &= \sum\nolimits_{s=1}^{S}\left(\hat{T}_{\mathrm{HT},s} - T\right)^2\Big/\sum\nolimits_{s=1}^{S}\left(\hat{T}_{\mathrm{GREG},s} - T\right)^2\end{aligned}$$

where $\hat{T}_s$ is an estimated total from sample s (either HT or GREG), $S = 10{,}000$ is the number of samples selected, and $\hat{T}_{\mathrm{HT},s}$ and $\hat{T}_{\mathrm{GREG},s}$ are the estimated HT and GREG totals from sample s. The empirical deff of an estimated total is computed as $\mathrm{empdeff}\left(\hat{T}\right) = S^{-1}\sum\nolimits_{s=1}^{S}\left(\hat{T}_s - \bar{\hat{T}}\right)^2\Big/\mathrm{Var}_{\mathrm{srswr}}\left(\hat{T}_{\mathrm{srswr}}\right)$ where $\bar{\hat{T}} = S^{-1}\sum\nolimits_{s=1}^{S}\hat{T}_s$ and $\mathrm{Var}_{\mathrm{srswr}}\left(\hat{T}_{\mathrm{srswr}}\right) \doteq N^2\sigma_y^2\big/n$.

The results for relbiases and MSEs are shown in Table 4.1. Both estimators of totals are approximately unbiased. The GREG is also more precise than the HT estimator, especially for Total Expenses, as evidenced by the MSE ratios larger than one.

Table 4.1
Simulation results of HT and GREG totals, 10,000 πps samples drawn from the SOI 2007 pseudopopulation EO data

	Variable of Interest					
	Total Liabilities (weakly correlated with X)			Total Expenses (strongly correlated with X)		
Estimates	$n = 100$	$n = 500$	$n = 1{,}000$	$n = 100$	$n = 500$	$n = 1{,}000$
Percent relbias(HT)	-0.13	0.07	0.03	-0.64	0.05	0.07
Percent relbias(GREG)	0.37	0.27	0.14	-0.12	-0.01	0.00
MSE ratio	1.17	1.20	1.19	34.89	50.11	48.26

Note A small number of samples were dropped in which either the matrix to be inverted for the GREG was singular or the GREG produced negative weights. The percentages of samples dropped were 3.6% for $n = 100,$ 1.2% for $n = 500,$ and 0.5% for $n = 1{,}000.$

We also computed the biases of the various estimated design effects across the 10,000 samples. The relbiases of the Kish, Spencer, and proposed design effect estimates are computed as

$$\text{relbias}\left(\text{deff}_K\right) = 100 \times \left(\overline{\text{deff}}_K - \text{edeff}\left(\hat{T}_{\text{HT}y}\right)\right)\Big/\text{edeff}\left(\hat{T}_{\text{HT}y}\right),$$

$$\text{relbias}\left(\text{deff}_S\right) = 100 \times \left(\overline{\text{deff}}_S - \text{edeff}\left(\hat{T}_{\text{HT}y}\right)\right)\Big/\text{edeff}\left(\hat{T}_{\text{HT}y}\right),$$

and

$$\text{relbias}\left(\text{deff}_H\right) = 100 \times \left(\overline{\text{deff}}_H - \text{edeff}\left(\hat{T}_{\text{GREG}}\right)\right)\Big/\text{edeff}\left(\hat{T}_{\text{GREG}}\right)$$

where $\overline{\text{deff}}_K,$ $\overline{\text{deff}}_S,$ and $\overline{\text{deff}}_H$ are the average Kish, Spencer, and proposed deff's over all samples. The terms $\text{edeff}\left(\hat{T}_{\text{HT}y}\right)$ and $\text{edeff}\left(\hat{T}_{\text{GREG}}\right)$ are computed in two ways: (1) as the simulation empdeff of $\hat{T}_{\text{HT}y}\left(\text{or } \hat{T}_{\text{GREG}}\right),$ and (2) as the average over all samples of the deff's of $\hat{T}_{\text{HT}y}$ computed from the `survey` package. The `survey` package's default method of estimating the deff from a particular sample uses a with-replacement variance estimate in the numerator. This corresponds to the sample design used to derive $\text{deff}_H.$ Results are displayed in Table 4.2.

For both variables of interest, we see large positive biases for the Kish design effect, and the design effects involving approximations. Thus, ignoring correlation components accounted for in the 'exact' Spencer and proposed design effects would lead to over-estimating the design effects.

The proposed estimator is closer to the `survey` package design effects than to the empirical simulation deff's of the GREG. Although the relbiases of deff_H are fairly large for Total Expenses when computed with respect to edeff, the empirical deff's themselves are small. We highlight the small magnitude of the Total Expenses (y_2) variable deff of 0.02 to put the relbiases into context. For example, the relbias of 12.9% for the exact version of our proposed estimator for $n = 500$ for y_2 corresponds to a difference in the third decimal place. Specifically, in this scenario, on average we over-estimate the deff by 0.003.

We can understand why calibration is more efficient for Expenses than for Liabilities by examining the distributions of y_i and u_i in one particular sample. Figures 4.2 and 4.3 show boxplots of u_i and y_i for each variable and sample size.

Table 4.2
Relative bias of design effect estimates, 10,000 πps samples drawn from the SOI 2007 pseudopopulation EO data

	Variable of Interest					
	Total Liabilities (weakly correlated with $\mathbf{X}$)			Total Expenses (strongly correlated with $\mathbf{X}$)		
	$n = 100$	$n = 500$	$n = 1{,}000$	$n = 100$	$n = 500$	$n = 1{,}000$
Empirical deff's[*]						
HT	0.51	0.50	0.50	0.56	0.65	0.64
GREG	0.43	0.42	0.42	0.02	0.02	0.02
Relative biases w.r.t. empirical deff's						
Kish[**]	158.7	158.3	158.3	132.8	101.7	104.7
Spencer[**]						
Exact	2.6	2.0	1.8	9.9	-4.5	-2.2
Zero-corr. approx.	96.1	98.0	98.4	91.2	70.1	73.7
Large $-N$ approx.	96.7	98.9	99.3	101.7	78.1	81.7
Proposed[***]						
Exact	-6.3	-1.6	0.2	25.3	12.9	8.1
Zero-corr. approx.	83.4	94.0	98.2	129.9	116.6	108.7
Relative biases w.r.t. average of `survey` *package deff's*						
Kish[**]	219.7	211.3	209.4	6,400.5	7,786.2	8,287.2
Spencer[**]						
Exact	3.1	0.8	0.5	3.5	-1.0	-1.5
Zero-corr. approx.	97.1	95.8	95.8	80.1	76.2	74.8
Large $-N$ approx.	97.7	96.7	96.7	90.0	84.5	82.8
Proposed[***]						
Exact	-0.9	-0.2	-0.1	11.3	-0.4	-0.1
Zero-corr. approx.	94.0	96.8	97.6	104.2	91.0	93.0

[*] Averages across the simulated samples;
[**] relative to the average of empirical HT deff's;
[***] relative to the average of empirical GREG deff's.

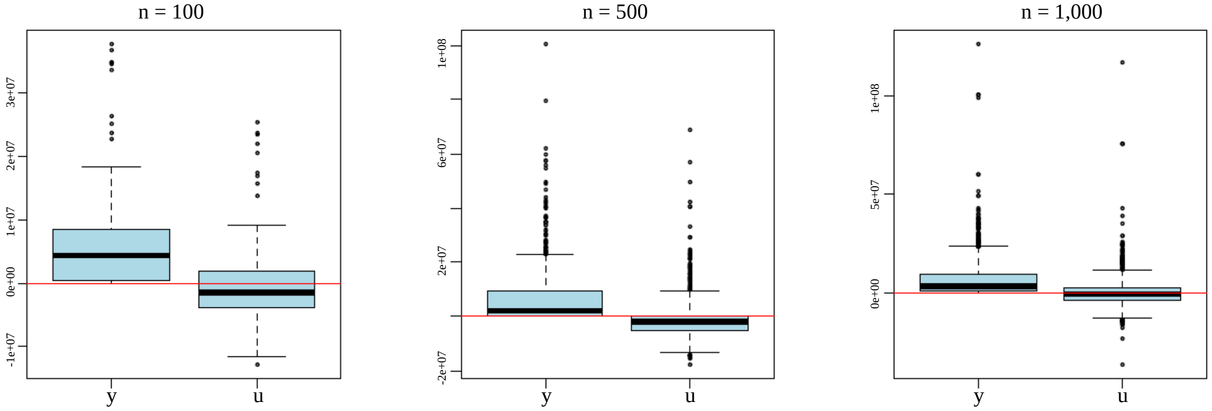


Figure 4.2 **Boxplots of y_i and $u_i -$values from ppswr samples from the 2007 SOI EO data, total liabilities variable (weakly correlated with $\mathbf{X}$).**

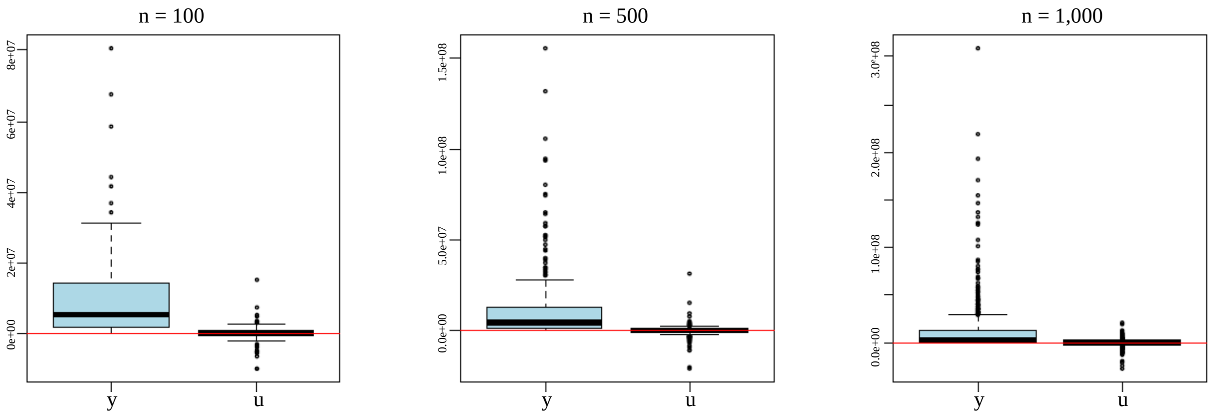


Figure 4.3 Boxplots of y_i and u_i-values from πps samples from the 2007 SOI EO data, total expenses variable (strongly correlated with $\mathbf{X}$).

The u_i-values in all of these samples have shorter ranges of values and less variation than y_i, particularly for the Total Expenses variable. This occurs since the Total Expenses variable is highly correlated with the calibration variable Total Revenue (see Figure 4.1) and explains why the direct and proposed design effect measures are so much smaller for Total Expenses.

4.2 Simulation study with a binary variable

The second simulation study illustrates the performance of the proposed estimator when estimating the total of a binary variable in a single-stage survey that uses poststratification.

We use the `nhis.large` population, which has $N = 21{,}588$ units, from the `PracTools` R package (Valliant, Dever and Kreuter 2015) to gauge the impact of poststratification weighting adjustments. The binary variable used is whether or not a person received Medicaid or not. Receipt of Medicaid, which is a social welfare program in the US, is an example of a variable that is collected in some telephone surveys. Missing values of Medicaid recipiency were recoded to be “no” responses. There is a fairly strong relationship between race-ethnicity, age, and whether Medicaid is received, as shown in Table 4.3 or Table 14.1 in Valliant, Dever and Kreuter (2013). The 15 age × race-ethnicity cells in the table will be used as poststrata, which is a typical procedure in telephone surveys.

Table 4.3
Population percentages of persons receiving medicaid, by age group and Hispanic status

	Hispanic Status		
Age Group	**Hispanic**	**Non-Hispanic White**	**Non-Hispanic Black or Other**
< 18 years	31.8	12.9	30.9
18-24	10.5	6.5	12.2
25-44	7.5	3.8	8.6
45-64	2.4	3.0	6.2
65+	26.8	3.7	16.2

In our simulation, we selected 10,000 simple random samples without replacement from the NHIS population. The HT estimator for the total number of persons receiving Medicaid is $N\bar{y}_s,$ where $\bar{y}_s$ is the proportion in sample s that receives Medicaid. Due to the relatively large number of poststrata and varying number of persons receiving Medicaid by poststratum, we include results only for samples of size $n = 500$ and $1{,}000$ since no collapsing of poststrata within a given particular sample was needed for these sample sizes.

The base weights for the $\mathrm{HT}-$estimator are simply $w_i = N/n.$ The variance of the poststratified estimator is 91% of that of $N\bar{y}_s$ in samples of $n = 500$ and 88% in samples of $n = 1{,}000.$ Since the base weights are constant, Spencer's design effects are not computable in this example. Therefore, only results for the Kish and proposed design effects are shown in Table 4.4.

Table 4.4
Relative bias of design effect estimates, 10,000 pps samples drawn from the NHIS pseudopopulation data

	Number of Persons Receiving Medicaid			
	$\boldsymbol{n = 500}$		$\boldsymbol{n = 1{,}000}$	
Empirical deff's[*]				
HT	0.97		0.95	
GREG	0.91		0.88	
	w.r.t. empirical deff	w.r.t. survey deff	w.r.t. empirical deff	w.r.t. survey deff
Relative biases (percent)				
Kish[**]	6.0	17.5	7.0	17.6
Proposed[***]				
Exact	-1.4	3.2	-0.9	5.0
Zero-corr. approx.	-1.5	2.9	-1.2	4.7

[*] Averages across the simulated samples;
[**] relative to the average of empirical HT deff's;
[***] relative to the average of empirical GREG deff's.

The Kish design effect has positive biases of 17.5% and 17.6% when computed with respect to the empirical deff's. The exact proposed design effects are positively biased with respect to the survey deff (3.2 and 5.0%), but much less so than the Kish estimator. In this example, the zero-correlation approximation is very similar to the exact version of the proposed estimator. The correlation components were negligible for these weighting adjustments within three decimal places.

5 Discussion, limitations, and conclusions

We propose a new design effect that gauges the impact of calibration weighting adjustments on an estimated total in single-stage sampling. Two existing design effects are the Kish (1965) "design effect due to weighting" and one due to Spencer (2000). Both of these are inadequate to reflect efficiency gains due to calibration. The Kish deff is a reasonable measure if equal weighting is optimal or nearly so, but does not reveal efficiencies that may accrue from sampling with varying probabilities. The Spencer deff does signal whether the HT (or pwr) estimator in varying probability sampling is more efficient than srs. But, the Spencer deff does not reflect any gains from using calibration.

The proposed design effect measures the impact of both sampling with varying probabilities and of using a calibration estimator, like the GREG, that takes advantage of auxiliary information. As we demonstrate empirically, the proposed design effects do not penalize unequal weights when the relationship between the survey variable and calibration covariate is strong. We also demonstrated empirically that the correlation components in the Spencer measure and our proposed measure can be important in some situations. It is not overly difficult to calculate these components, and these should be incorporated when possible to avoid over estimates of the design effects. However, the high correlations between survey and auxiliary variables that we observed in our establishment pseudopopulation data may be unattainable for some surveys that lack auxiliary information. In cases where the auxiliary information is ineffective or is not used, the proposed measure approximates Kish's deff. The measure presented here is applicable to single-stage sampling but can be extended to more complex sample designs, like cluster sampling.

Our measure uses the model underlying the general regression estimator to extend the Spencer measure. The survey variable, covariates, and weights are required to produce the design effect estimate. Since the variance (3.2) is approximately correct in large samples for all calibration estimators, our design effect should reflect the effects of many forms of commonly used weighting adjustment methods, including poststratification, raking, and the GREG estimator. Although design effects that do account for these adjustments can be computed directly from estimated variances, it is important for practitioners to understand that the existing Kish and Spencer deff's do not reflect any gains from those adjustments. The deff introduced in this paper, thus, serves as a corrective to that deficiency.

For practical consideration, the deff in (3.4) is available in the `deffH` function in the R `PracTools` package; see Valliant et al. (2015) for documentation and examples.

Acknowledgements

We thank the referees for their thorough reviews which improved the presentation. Any opinions expressed are those of the authors and do not reflect those of the Internal Revenue Service.

Appendix

Proposed design effect in single-stage sampling

The appendix sketches the derivation of the proposed deff. Most notation was defined in the previous sections of the paper. The average population one-draw probability is $\bar{P} = N^{-1}\sum_{i=1}^{N} p_i$. Assume that the design satisfies $\bar{P} = N^{-1}$. Consider the model $y_i = \alpha + \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i$. If the full finite population were available, then the least-squares population regression line would be

$$y_i = A + \mathbf{x}_i^T \mathbf{B} + e_i, \tag{A.1}$$

where A and $\mathbf{B}$ are the values found by fitting an ordinary least squares regression line in the full finite population. That is, $A = \bar{Y} - \mathbf{B}\bar{\mathbf{X}}$, $\mathbf{B} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{y}$, where $\mathbf{X}$ is the $N \times p$ population matrix of auxiliary variables, $\bar{Y} = N^{-1} \sum_{i=1}^{N} y_i$ is the population mean, and $\bar{\mathbf{X}}$ is the vector of population means of the x's. The e_i's are defined as the finite population residuals, $e_i = y_i - A - \mathbf{x}_i^T \mathbf{B}$, and are not superpopulation model errors. Denote the population variance of the y's, e's, e^2, and weights as $\sigma_y^2, \sigma_e^2, \sigma_{e^2}^2, \sigma_w^2$, e.g., $\sigma_y^2 = N^{-1} \sum_{i=1}^{N} (y_i - \bar{Y})^2$, and the finite population correlations between the variables in the subscripts as ρ_{yp}, ρ_{ew}, and $\rho_{e^2 w}$. The GREG theoretical design-variance in with-replacement sampling is

$$\begin{aligned} \operatorname{Var}\left(\hat{T}_{\text{GREG}}\right) &= n^{-1} \sum_{i=1}^{N} p_i \left(e_i / p_i - E_U\right)^2 \\ &= n^{-1} \left(\sum_{i=1}^{N} e_i^2 / p_i - E_U^2\right), \end{aligned} \tag{A.2}$$

where $E_U = \sum_{i=1}^{N} e_i$. Using the model in (A.1) produces a design effect with several complex terms, many of which contain correlations that cannot be dropped as in Spencer's approximation. The design effect can be simplified using an alternative formulation: $u_i = A + e_i$, where $u_i = y_i - \mathbf{x}_i^T \mathbf{B}$. First, we rewrite the population total of the e_i's as $E_U = \sum_{i=1}^{N} e_i = N\bar{U} - NA$, where $\bar{U} = N^{-1} \sum_{i=1}^{N} u_i$. From this, $E_U^2 = (N\bar{U})^2 + (NA)^2 - 2N^2 \bar{U} A$. Second, using $w_i = (np_i)^{-1}$, or $p_i = (nw_i)^{-1}$, we rewrite the component $\sum_{i=1}^{N} e_i^2 / p_i$ as

$$\begin{aligned} \sum_{i=1}^{N} e_i^2 / p_i &= \sum_{i=1}^{N} \frac{(u_i - A)^2}{(nw_i)^{-1}} \\ &= n \sum_{i=1}^{N} w_i u_i^2 + nA^2 \sum_{i=1}^{N} w_i - 2nA \sum_{i=1}^{N} w_i u_i. \end{aligned} \tag{A.3}$$

Subtracting E_U^2 from (A.3) and dividing by n gives

$$\begin{aligned} n^{-1} \left(\sum_{i=1}^{N} e_i^2 / p_i - E_U^2\right) &= \sum_{i=1}^{N} w_i u_i^2 - n^{-1} (N\bar{U})^2 \\ &\quad + A^2 \left(\sum_{i=1}^{N} w_i - n^{-1} N^2\right) \\ &\quad + n^{-1} 2N^2 \bar{U} A - 2A \sum_{i=1}^{N} w_i u_i. \end{aligned} \tag{A.4}$$

Following Spencer's approach using the covariance substitutions, the first and fifth terms in (A.4) can be rewritten as $\sum_{i=1}^{N} w_i u_i^2 = N\rho_{u^2 w} \sigma_{u^2} \sigma_w + N\bar{W} \left(\sigma_u^2 + \bar{U}^2\right)$ and $\sum_{i=1}^{N} w_i u_i = N\rho_{uw} \sigma_u \sigma_w + N\bar{W}\bar{U}$. Plugging these back into the variance (A.4) gives

$$\begin{aligned} n^{-1} \left(\sum_{i=1}^{N} e_i^2 / p_i - E_U^2\right) &= N\rho_{u^2 w} \sigma_{u^2} \sigma_w + N\bar{W} \left(\sigma_u^2 + \bar{U}^2\right) - n^{-1} (N\bar{U})^2 \\ &\quad + NA^2 \left(\bar{W} - n^{-1} N\right) \\ &\quad + 2n^{-1} N^2 \bar{U} A - 2A \left(N\rho_{uw} \sigma_u \sigma_w + N\bar{W}\bar{U}\right). \end{aligned} \tag{A.5}$$

The variance of the $\text{pwr}-$estimator under simple random sampling with replacement, where $p_i = N^{-1}$, reduces to $\text{Var}_{\text{srswr}}\left(\hat{T}_{\text{pwr}}\right) = N^2\sigma_y^2/n$. Taking the ratio of (A.5) to the $\text{pwr}-$variance gives the following design effect:

$$\begin{aligned}\text{Deff}_H &= \text{Var}_{\text{GREG}}\left(\hat{T}_{\text{cal}}\right)\Big/\text{Var}_{\text{srswr}}\left(\hat{T}_{\text{pwr}}\right)\\ &= \frac{n\bar{W}}{N}\left(\frac{\sigma_u^2}{\sigma_y^2}\right) + \frac{\left(\bar{U}-A\right)^2}{\sigma_y^2}\left(\frac{n\bar{W}}{N}-1\right) \qquad \text{(A.6)}\\ &+ \frac{n\sigma_w}{N\sigma_y^2}\left(\rho_{u^2w}\sigma_{u^2} - 2A\rho_{uw}\sigma_u\right).\end{aligned}$$

Since $u_i = A + e_i, \bar{U} = A,$ (A.6) becomes

$$\text{Deff}_H = \frac{n\bar{W}}{N}\left(\frac{\sigma_u^2}{\sigma_y^2}\right) + \frac{n\sigma_w}{N\sigma_y^2}\left(\rho_{u^2w}\sigma_{u^2} - 2A\rho_{uw}\sigma_u\right). \qquad \text{(A.7)}$$

We estimate measure (A.7) with

$$\text{deff}_H \approx \left(1 + [\text{CV}(\mathbf{w})]^2\right)\frac{\hat{\sigma}_u^2}{\hat{\sigma}_y^2} + \frac{n\hat{\sigma}_w}{N\hat{\sigma}_y^2}\left(\hat{\rho}_{u^2w}\hat{\sigma}_{u^2} - 2\hat{\alpha}\hat{\rho}_{uw}\hat{\sigma}_u\right), \qquad \text{(A.8)}$$

where the model parameter estimates are defined in Sections 2.3 and 3.

References

Brick, M., and Montaquila, J. (2009). Nonresponse. In *Handbook of Statistics, Sample Surveys: Design, Methods and Application,* (Eds., D. Pfeffermann and C.R. Rao), 29A, Amsterdam: Elsevier BV.

Chambers, J.M., Cleveland, W.S., Kleiner, B. and Tukey, P.A. (1983). *Graphical Methods for Data Analysis*. Pacific Grove CA: Wadsworth.

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

Deville, J.-C., Särndal, C.-E. and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.

Hansen, M.H., and Hurwitz, W.N. (1943). On the theory of sampling from a finite population. *Annals of Mathematical Statistics*, 14, 333-362.

Horvitz, D., and Thompson, D. (1952). A generalisation of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

Kalton, G., and Flores-Cervantes, A. (2003). Weighting methods. *Journal of Official Statistics*, 19 (2), 81-97.

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

Kish, L. (1990). Weighting: Why, when, and how? *Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods,* American Statistical Association, 121-129.

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

Kott, P. (2009). Calibration weighting: Combining probability samples and linear prediction models. In *Handbook of Statistics, Sample Surveys: Design, Methods and Application,* (Eds., D. Pfeffermann and C.R. Rao), 29B, Amsterdam: Elsevier BV.

Lumley, T. (2012). Survey: Analysis of complex survey samples. R package version 3.28-2.

Park, I., and Lee, H. (2004). Design effects for the weighted mean and total estimators under complex survey sampling. *Survey Methodology*, 30, 2, 183-193.

Rao, J.N.K., and Scott, A.J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *Annals of Statistics*, 12, 46-60.

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

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

Spencer, B.D. (2000). An approximate design effect for unequal weighting when measurements may correlate with selection probabilities. *Survey Methodology*, 26, 2, 137-138.

Statistics of Income (2011). 2007 Charities & Tax-Exempt Microdata Files. Available at: http://www.irs.gov/uac/SOI-Tax-Stats-2007-Charities-&-Tax-Exempt-Microdata-Files.

Valliant, R., Dever, J.A. and Kreuter, F. (2013). *Practical Tools for Designing and Weighting Survey Samples*. New York: Springer.

Valliant, R., Dever, J.A. and Kreuter, F. (2015). PracTools: Tools for Designing and Weighting Survey Samples. R package version 0.2. http://CRAN.R-project.org/package=PracTools.

Model-based small area estimation under informative sampling

François Verret, J.N.K. Rao and Michael A. Hidiroglou[1]

Abstract

Unit level population models are often used in model-based small area estimation of totals and means, but the models may not hold for the sample if the sampling design is informative for the model. As a result, standard methods, assuming that the model holds for the sample, can lead to biased estimators. We study alternative methods that use a suitable function of the unit selection probability as an additional auxiliary variable in the sample model. We report the results of a simulation study on the bias and mean squared error (MSE) of the proposed estimators of small area means and on the relative bias of the associated MSE estimators, using informative sampling schemes to generate the samples. Alternative methods, based on modeling the conditional expectation of the design weight as a function of the model covariates and the response, are also included in the simulation study.

Key Words: Augmented model; Empirical best linear unbiased prediction (EBLUP); Nested error model; Pseudo-EBLUP.

1 Introduction

Estimates of population totals and means are often required for small subpopulations (or areas). Traditional area-specific direct estimators are not reliable if the area sample size is small. As a result, it becomes necessary to "borrow strength" across areas through indirect estimation based on models that provide a link to related areas. Linking models make use of auxiliary population information either at the area level or at the unit level. Rao (2003, Chapter 7) gives a detailed account of area level and unit level models that are widely used for small area estimation.

Suppose that the population of interest, U, consists of M non-overlapping areas with N_i elements in the i^{th} area $(i = 1, \ldots, M)$. A sample, s, of m areas is first selected using a specified sampling scheme with inclusion probabilities $\pi_i = mp_i \, (i = 1, \ldots, M)$, where p_i denotes the selection probability of area i. Subsamples s_i of specified sizes n_i are then independently selected from the sampled areas i according to specified sampling schemes with selection probabilities $p_{j|i} \left(\sum_{j=1}^{N_i} p_{j|i} = 1 \right)$ such that the second-stage inclusion probabilities are $\pi_{j|i} = n_i p_{j|i}$ for unit j in area $i \; (j = 1, \ldots, N_i)$. Typically, the selection probability $p_{j|i} = b_{ij} \Big/ \sum_{k=1}^{N_i} b_{ik}$, where b_{ij} is a size measure related to the response variable y_{ij}. In this paper, we focus on the special case where all the areas are sampled, $m = M$.

We assume a nested error linear regression model for the population, based on covariates $\mathbf{x}_{ij}$ related to the response variable y_{ij}. The population model is assumed to be given by

[1] François Verret, Statistics Canada, 23 B, R.-H.-Coats Building, Ottawa, Ontario, Canada, K1A 0T6. E-mail: francois.verret@canada.ca; J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada, K1S 5B6. E-mail: jrao@math.carleton.ca; Michael A. Hidiroglou, Statistics Canada, 16 D, R.-H.-Coats Building, Ottawa, Ontario, Canada, K1A 0T6. E-mail: mike.hidiroglou@canada.ca.

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + v_i + e_{ij};\ j = 1, \ldots, N_i;\ i = 1, \ldots, M, \tag{1.1}$$

where $v_i \overset{\text{iid}}{\sim} N(0, \sigma_v^2)$ are random small area effects that are independent of the unit-level errors $e_{ij} \overset{\text{iid}}{\sim} N(0, \sigma_e^2), \mathbf{x}_{ij} = (1, x_{ij1}, x_{ij2}, \ldots, x_{ijp})^T$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$. Parameters of interest are the small area means $\bar{Y}_i = N_i^{-1} \sum_{j=1}^{N_i} y_{ij}$ which may be approximated by $\mu_i = \bar{\mathbf{X}}_i^T \beta + v_i$, if the area sizes N_i are large, where $\bar{\mathbf{X}}_i = N_i^{-1} \sum_{j=1}^{N_i} \mathbf{x}_{ij}$ is the known population mean of $\mathbf{x}$ for area i.

Efficient model-based estimators of the area means μ_i may be obtained if the sampling design is non-informative for the model, which implies that the sample and the population models coincide. In particular, empirical best linear unbiased prediction (EBLUP) estimators (Henderson 1975), based on the assumed sample model under non-informative sampling, may be used to estimate small area means $\bar{Y}_i$ or μ_i (see Section 2 and Rao 2003, Chapter 7). However, in many practical situations the selection probabilities $p_{j|i}$ may be related to the associated y_{ij} even after conditioning on the covariates $\mathbf{x}_{ij}$. In such cases, we have “informative sampling” in the sense that the population model (1.1) no longer holds for the sample. For example, Pfeffermann and Sverchkov (2007) assumed that the sampled unit design weight $w_{j|i} = \pi_{j|i}^{-1}$ is random with conditional expectation

$$\begin{aligned} E_{s_i}\left(w_{j|i} \middle| \mathbf{x}_{ij}, y_{ij}, v_i\right) &= E_{s_i}\left(w_{j|i} \middle| \mathbf{x}_{ij}, y_{ij}\right) \\ &= k_i \exp\left(\mathbf{x}_{ij}^T \mathbf{a} + b y_{ij}\right), \end{aligned} \tag{1.2}$$

where $\mathbf{a}$ and b are fixed unknown constants and

$$k_i = N_i n_i^{-1} \left\{ \sum_{j=1}^{N_i} \exp\left(-\mathbf{x}_{ij}^T \mathbf{a} - b y_{ij}\right) \Big/ N_i \right\}.$$

Under informative sampling within areas, the EBLUP estimator of $\bar{Y}_i$, assuming that model (1.1) holds for the sample, may be heavily biased. It is, therefore, necessary to develop estimators that can account for sample selection bias and thus reduce estimation bias. Pfeffermann and Sverchkov (2007) developed a bias-adjusted estimator of the mean $\bar{Y}_i$ under the assumption (1.2) on the design weights $w_{j|i}$ and assuming that the sample model is a nested error model

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\alpha} + u_i + h_{ij};\ j = 1, \ldots, n_i;\ i = 1, \ldots, M, \tag{1.3}$$

where $u_i \overset{\text{iid}}{\sim} N(0, \sigma_u^2)$, and $h_{ij} \big| j \in s_i \overset{\text{iid}}{\sim} N(0, \sigma_h^2)$. Pfeffermann and Sverchkov (2007) noted that under a sampling scheme satisfying (1.2) the population model is also a nested error model but with different parameters. However, they do not use the form of the population model. The sample model (1.3) is identified after fitting the model to the sample data and then doing some model diagnostics. Similarly, model (1.2) on the weights is identified from the sample data $\{w_{j|i}, y_{ij}, \mathbf{x}_{ij}, j \in s_i, i \in s\}$. Their estimators are noted (PS) in the following.

Prasad and Rao (1999) and You and Rao (2002) developed pseudo-EBLUP estimators of small area means μ_i that depend on the sampling weights $w_{j|i}$, assuming non-informative sampling for the model

(1.1). Their motivation for pseudo-EBLUP is to ensure design consistency as the area sample size, n_i, increases. The estimators of You and Rao (note (YR) in the following) also satisfy a benchmarking property in the sense that the associated estimators of area totals add up to a reliable direct estimator of the total, unlike the EBLUP estimators. Stefan (2005) studied the empirical performance of pseudo-EBLUP estimators under informative sampling for model (1.1) and showed that the pseudo-EBLUP leads to smaller bias compared to the EBLUP.

The main purpose of our paper is to study augmented sample models of the form

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta}_0 + g\left(p_{j|i}\right)\delta_0 + \tilde{v}_i + \tilde{e}_{ij};\ j = 1, \ldots, n_i;\ i = 1, \ldots, M \tag{1.4}$$

for a suitably defined function $g_{ij} = g\left(p_{j|i}\right)$ of the probability $p_{j|i}$, where $\tilde{v}_i \overset{\text{iid}}{\sim} N\left(0, \sigma_{v0}^2\right)$ and independent of $\tilde{e}_{ij} \overset{\text{iid}}{\sim} N\left(0, \sigma_{e0}^2\right)$, and $\boldsymbol{\beta}_0 = \left(\beta_{00}, \beta_{01}, \ldots, \beta_{0p}\right)^T$. The sample model (1.4) is identified after fitting the model to sample data for different choices of the function $g(\cdot)$ and checking their adequacy. For example, residuals r_{ij} from fitting the model (1.4) without the augmenting variable $g\left(p_{j|i}\right)$ may be plotted against $g\left(p_{j|i}\right)$ to select $g(\cdot)$. The identified augmented sample model will also hold for the population (Skinner 1994, Rao 2003, Section 5.3). Possible choices of $g\left(p_{j|i}\right)$ are $p_{j|i}, \log p_{j|i}, w_{j|i}$ and $n_i w_{j|i} = p_{j|i}^{-1}$.

From the augmented sample model (1.4) we obtain the EBLUP estimators of $\bar{Y}_i$ or $\mu_i = \bar{\mathbf{X}}_i^T \boldsymbol{\beta}_0 + \bar{G}_i \delta_0 + \tilde{v}_i$, the approximate area mean under the augmented population model, where $\bar{G}_i$ is the area mean of the population values $g\left(p_{j|i}\right) \equiv g_{ij}$. The EBLUP of $\bar{Y}_i$ or μ_i requires the knowledge of $\bar{G}_i$ which depends on all the population values $p_{j|i}$. However, the choice $g\left(p_{j|i}\right) = p_{j|i}$ gives $\bar{G}_i = 1/N_i$ and the choice $g\left(p_{j|i}\right) = n_i w_{j|i}$ gives $\bar{G}_i = n_i \bar{W}_i$, where $\bar{W}_i$ is the area population mean of the weights $w_{j|i}$. The means $\bar{W}_i$ are often known in practice. Pseudo-EBLUP estimators under the augmented model are also studied.

We conducted a simulation study under the design-model (or pm) framework to study the bias and MSE of the proposed estimators relative to EBLUP and pseudo-EBLUP estimators based on non-informative sampling, and the bias-adjusted estimators of Pfeffermann and Sverchkov (2007). We also studied the performance of MSE estimators in terms of relative bias.

Section 2 summarizes the existing model-based methods for estimating the small area means $\bar{Y}_i$ or μ_i. Proposed methods based on the augmented sample model (1.4) are presented in Section 3. The results of the simulation study are reported in Section 4. Concluding remarks are given in Section 5.

2 Existing methods

2.1 Estimators of small area means

Suppose that the population model (1.1) holds for the sample. Then the EBLUP estimator of $\mu_i = \bar{\mathbf{X}}_i^T \boldsymbol{\beta} + v_i$ is given by

$$\hat{\mu}_i^H = \bar{\mathbf{X}}_i^T \hat{\boldsymbol{\beta}} + \hat{v}_i = \hat{\gamma}_i \bar{y}_i + (\bar{\mathbf{X}}_i - \hat{\gamma}_i \bar{\mathbf{x}}_i)^T \hat{\boldsymbol{\beta}}, \tag{2.1}$$

where $\hat{\gamma}_i = \hat{\sigma}_v^2 \big/ \left(\hat{\sigma}_v^2 + \hat{\sigma}_e^2 / n_i\right), \bar{y}_i = \sum_{j=1}^{n_i} y_{ij} \big/ n_i, \bar{\mathbf{x}}_i = \sum_{j=1}^{n_i} \mathbf{x}_{ij} \big/ n_i$ are the unweighted sample means of the response variable y and the covariates $\mathbf{x}$ and $\hat{v}_i = \hat{\gamma}_i \left(\bar{y}_i - \bar{\mathbf{x}}_i^T \hat{\boldsymbol{\beta}}\right)$. Further,

$$\hat{\boldsymbol{\beta}} = \left\{ \sum_{i=1}^{M} \sum_{j=1}^{n_i} \mathbf{x}_{ij} \left(\mathbf{x}_{ij} - \hat{\gamma}_i \bar{\mathbf{x}}_i\right)^T \right\}^{-1} \left\{ \sum_{i=1}^{M} \sum_{j=1}^{n_i} \left(\mathbf{x}_{ij} - \hat{\gamma}_i \bar{\mathbf{x}}_i\right) y_{ij} \right\} \tag{2.2}$$

and $\hat{\sigma}_e^2$ and $\hat{\sigma}_v^2$ are obtained by the method of fitting of constants (Battese, Harter and Fuller (1988); Rao (2003, Chapter 7)) or restricted maximum likelihood (REML). The EBLUP estimator of the area mean $\bar{Y}_i$ may be written in terms of $\hat{\mu}_i^H$ as

$$\hat{\bar{Y}}_i^H = N_i^{-1} \left[(N_i - n_i) \hat{\mu}_i^H + n_i \left\{ \bar{y}_i + (\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_i)^T \hat{\boldsymbol{\beta}} \right\} \right], \tag{2.3}$$

(see Rao 2003, page 141). Note that $\hat{\bar{Y}}_i^H \approx \hat{\mu}_i^H$ if the sampling fraction n_i / N_i is sufficiently small. The EBLUP estimator $\hat{\bar{Y}}_i^H$ is design consistent under simple random sampling (SRS) or stratified SRS with proportional allocation within area i, leading to equal $\pi_{j|i}$.

The pseudo-EBLUP estimator of μ_i is given by

$$\hat{\mu}_i^{\text{YR}} = \hat{\gamma}_{iw} \bar{y}_{iw} + (\bar{\mathbf{X}}_i - \hat{\gamma}_{iw} \bar{\mathbf{x}}_{iw})^T \hat{\boldsymbol{\beta}}_w, \tag{2.4}$$

where we denote by $\tilde{w}_{j|i} = w_{j|i} \big/ \sum_{k=1}^{n_i} w_{k|i}$ the normalized weights, $\hat{\gamma}_{iw} = \hat{\sigma}_v^2 \big/ \left(\hat{\sigma}_v^2 + \delta_i^2 \hat{\sigma}_e^2\right)$ with $\delta_i^2 = \sum_{j=1}^{n_i} \tilde{w}_{j|i}^2, \bar{y}_{iw} = \sum_{j=1}^{n_i} \tilde{w}_{j|i} y_{ij}, \bar{\mathbf{x}}_{iw} = \sum_{j=1}^{n_i} \tilde{w}_{j|i} \mathbf{x}_{ij}$ are the i^{th} area weighted means of y and $\mathbf{x}$, and

$$\hat{\boldsymbol{\beta}}_w = \left[\sum_{i=1}^{M} \sum_{j=1}^{n_i} w_{j|i} \mathbf{x}_{ij} \left(\mathbf{x}_{ij} - \hat{\gamma}_{iw} \bar{\mathbf{x}}_{iw}\right)^T \right]^{-1} \left[\sum_{i=1}^{M} \sum_{j=1}^{n_i} w_{j|i} \left(\mathbf{x}_{ij} - \hat{\gamma}_{iw} \bar{\mathbf{x}}_{iw}\right) y_{ij} \right]. \tag{2.5}$$

The pseudo-EBLUP estimator $\hat{\mu}_i^{\text{YR}}$ is design consistent under arbitrary selection probabilities $p_{j|i}$ unlike the EBLUP $\hat{\bar{Y}}_i^H$.

Pfeffermann and Sverchkov (2007) studied estimation of small area means under informative sampling, assuming model (1.3) for the sample data and model (1.2) for the weights $w_{j|i}$. Under this assumption, they obtained an estimator of $\bar{Y}_i$ that provides protection against informative sampling. It is given by

$$\hat{\bar{Y}}_i^{\text{PS}} = N_i^{-1} \left[(N_i - n_i) \hat{\mu}_{iu}^H + n_i \left\{ \bar{y}_i + (\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_i)^T \hat{\boldsymbol{\alpha}} \right\} + (N_i - n_i) \hat{b} \hat{\sigma}_h^2 \right], \tag{2.6}$$

where $\hat{\mu}_{iu}^H = \bar{\mathbf{X}}_i^T \hat{\boldsymbol{\alpha}} + \hat{u}_i$ is the EBLUP estimator of $\mu_{iu} = \bar{\mathbf{X}}_i^T \boldsymbol{\alpha} + u_i$ under the sample model (1.3) and $\hat{b}$ is an estimator of b in the model (1.2) for the weights $w_{j|i}$. Note that $\left(\hat{\boldsymbol{\alpha}}, \hat{u}_i, \hat{\sigma}_u^2, \hat{\sigma}_h^2\right)$ is identical to $\left(\hat{\boldsymbol{\beta}}, \hat{v}_i, \hat{\sigma}_v^2, \hat{\sigma}_e^2\right)$ obtained by assuming that the population model (1.1) holds for the sample. Therefore, we can also express $\hat{\bar{Y}}_i^{\text{PS}}$ as

$$\hat{\hat{Y}}_i^{\text{PS}} = \hat{\hat{Y}}_i^{H} + \left(1 - \frac{n_i}{N_i}\right)\hat{b}\hat{\sigma}_e^2, \tag{2.7}$$

noting that $\hat{\mu}_i^H = \hat{\mu}_{iu}^H$.

The last term in (2.7) corrects for any bias due to informative sampling under (1.2). PS obtained the estimator $\hat{b}$ of b in (2.6) by regressing the sampling weights $w_{j|i}$ on $k_i \exp\left(\mathbf{x}_{ij}^T\mathbf{a} + by_{ij}\right)$. The coefficients $k_i, \mathbf{a}$ and b may be estimated by fitting the model (1.2) using procedure NLIN in SAS or function nls in Splus. This involves iterative calculations and the initial values for $\mathbf{a}$ and b are obtained by regressing $\log\left(w_{j|i}\right)$ on $\mathbf{x}_{ij}$ and y_{ij}. Initial values for $k_i, i = 1, \ldots, M$ are taken as $k_i = N_i / n_i$.

2.2 MSE estimation

The mean squared error (MSE) of the EBLUP estimator $\hat{\mu}_i^H$, assuming non-informative sampling, is estimated by

$$\text{mse}\left(\hat{\mu}_i^H\right) = g_{1i}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right) + g_{2i}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right) + 2g_{3i}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right), \tag{2.8}$$

where

$$\begin{aligned}
g_{1i}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right) &= (1 - \hat{\gamma}_i)\,\hat{\sigma}_v^2, g_{2i}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right) = (\bar{\mathbf{X}}_i - \hat{\gamma}_i\bar{\mathbf{x}}_i)^T \left(\sum_{i=1}^M \mathbf{X}_i^T \hat{\mathbf{V}}_i^{-1}\mathbf{X}_i\right)^{-1} (\bar{\mathbf{X}}_i - \hat{\gamma}_i\bar{\mathbf{x}}_i), \\
g_{3i}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right) &= \hat{\gamma}_i\,(1 - \hat{\gamma}_i)^2\,\hat{\sigma}_e^{-4}\hat{\sigma}_v^{-2} h\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right), \\
h\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right) &= \hat{\sigma}_e^4 \,\text{var}\left(\hat{\sigma}_v^2\right) - 2\hat{\sigma}_e^2\hat{\sigma}_v^2 \,\text{cov}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right) + \hat{\sigma}_v^4 \,\text{var}\left(\hat{\sigma}_e^2\right),
\end{aligned}$$

$\hat{\mathbf{V}}_i = \hat{\sigma}_e^2\mathbf{I}_{n_i} + \hat{\sigma}_v^2\mathbf{1}_{n_i}\mathbf{1}_{n_i}^T$ and $\mathbf{X}_i^T = \left(\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i}\right)$. The matrix $\sum_{i=1}^M \mathbf{X}_i^T \hat{\mathbf{V}}_i^{-1}\mathbf{X}_i$ may be expressed explicitly as $\hat{\sigma}_e^{-2}\sum_{i=1}^M\sum_{j=1}^{n_i}\left(\mathbf{x}_{ij}\mathbf{x}_{ij}^T - \hat{\gamma}_i\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^T\right)$. The MSE estimator (2.8) is unbiased to second order under non-informative sampling (Rao 2003, Chapter 7). We refer the reader to Rao (2003, page 142) for the corresponding MSE estimator of $\hat{\hat{Y}}_i^H$.

The MSE of the pseudo-EBLUP estimator $\hat{\mu}_i^{\text{YR}}$ is estimated by

$$\text{mse}\left(\hat{\mu}_i^{\text{YR}}\right) = g_{1iw}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right) + g_{2iw}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right) + 2g_{3iw}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right), \tag{2.9}$$

where

$$g_{1iw}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right) = (1 - \hat{\gamma}_{iw})\,\hat{\sigma}_v^2, g_{2iw}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right) = (\bar{\mathbf{X}}_i - \hat{\gamma}_{iw}\bar{\mathbf{x}}_{iw})^T\,\mathbf{\Phi}_w\,(\bar{\mathbf{X}}_i - \hat{\gamma}_{iw}\bar{\mathbf{x}}_{iw}),$$

$$\begin{aligned}
\mathbf{\Phi}_w &= \hat{\sigma}_e^2\left(\sum_{i=1}^M\sum_{j=1}^{n_i}\mathbf{x}_{ij}\mathbf{z}_{ij}^T\right)^{-1}\left(\sum_{i=1}^M\sum_{j=1}^{n_i}\mathbf{z}_{ij}\mathbf{z}_{ij}^T\right)\left\{\left(\sum_{i=1}^M\sum_{j=1}^{n_i}\mathbf{x}_{ij}\mathbf{z}_{ij}^T\right)^{-1}\right\}^T \\
&+ \hat{\sigma}_v^2\left(\sum_{i=1}^M\sum_{j=1}^{n_i}\mathbf{x}_{ij}\mathbf{z}_{ij}^T\right)^{-1}\left\{\sum_{i=1}^M\left(\sum_{j=1}^{n_i}\mathbf{z}_{ij}\right)\left(\sum_{j=1}^{n_i}\mathbf{z}_{ij}\right)^T\right\}\left\{\left(\sum_{i=1}^M\sum_{j=1}^{n_i}\mathbf{x}_{ij}\mathbf{z}_{ij}^T\right)^{-1}\right\}^T,
\end{aligned}$$

$$g_{3iw}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right) = \hat{\gamma}_{iw}\,(1 - \hat{\gamma}_{iw})^2\,\hat{\sigma}_e^{-4}\hat{\sigma}_v^{-2} h\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right)$$

and $\mathbf{z}_{ij} = w_{ij}\left(\mathbf{x}_{ij} - \hat{\gamma}_{iw}\bar{\mathbf{x}}_{iw}\right)$; see You and Rao (2002). The MSE estimator (2.9) is obtained by ignoring a cross-product term in $\mathrm{MSE}\left(\hat{\mu}_i^{\mathrm{YR}}\right)$. Torabi and Rao (2010) obtained a MSE estimator that accounts for the missing cross-product term and that is unbiased to second order under non-informative sampling. However, it is computationally more intensive than (2.9). It was not used in the simulation study (Section 4) since it would have slowed down the simulations significantly. A few simulation trials, however, revealed that the two MSE estimators give similar results under the simulation set-up used in Section 4.

Pfeffermann and Sverchkov (2007) proposed a parametric bootstrap method to estimate the MSE of the bias-adjusted estimator $\hat{\bar{Y}}_i^{\mathrm{PS}}$ given by (2.6). We have not included this MSE estimator in our simulation study.

3 Proposed method

The proposed method of estimating the small area means, $\bar{Y}_i$, is simple. It uses the standard EBLUP estimator under the augmented sample model (1.4). The model parameters $\left(\sigma_{v0}^2, \sigma_{e0}^2\right)$ and $\left(\boldsymbol{\beta}_0, \delta_0\right)$ are estimated by REML and weighted least squares (WLS) respectively. The EBLUP estimator of μ_i under the augmented model (1.4) is given by

$$\hat{\mu}_{i(a)}^{H} = \hat{\gamma}_{i0}\bar{y}_i + (\bar{\mathbf{X}}_i - \hat{\gamma}_{i0}\bar{\mathbf{x}}_i)^T\,\hat{\boldsymbol{\beta}}_0 + (\bar{G}_i - \hat{\gamma}_{i0}\bar{g}_i)\,\hat{\delta}_0, \tag{3.1}$$

where $\hat{\gamma}_{i0} = \hat{\sigma}_{v0}^2 \big/ \left(\hat{\sigma}_{v0}^2 + \hat{\sigma}_{e0}^2 / n_i\right), \left(\hat{\boldsymbol{\beta}}_0^T, \hat{\delta}_0\right)$ is the WLS estimator of $\left(\boldsymbol{\beta}_0^T, \delta_0\right)$ and $\bar{g}_i = \sum_{j=1}^{n_i} g_{ij} \big/ n_i$. Note that $\hat{\mu}_{i(a)}^{H}$ assumes that $\bar{G}_i$ is known. The EBLUP estimator of $\bar{Y}_i$ under the augmented model may be written in terms of $\hat{\mu}_{i(a)}^{\mathrm{H}}$ as

$$\hat{\bar{Y}}_{i(a)}^{H} = N_i^{-1}\left[(N_i - n_i)\,\hat{\mu}_{i(a)}^{\mathrm{H}} + n_i\left\{\bar{y}_i + (\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_i)^T\,\hat{\boldsymbol{\beta}}_0 + (\bar{G}_i - \bar{g}_i)\,\hat{\delta}_0\right\}\right]. \tag{3.2}$$

The pseudo-EBLUP estimator of μ_i under the augmented model (1.4) is similarly obtained by modifying (3.1) as

$$\hat{\mu}_{i(a)}^{\mathrm{YR}} = \hat{\gamma}_{i0w}\bar{y}_{iw} + (\bar{\mathbf{X}}_i - \hat{\gamma}_{i0w}\bar{\mathbf{x}}_{iw})^T\,\hat{\boldsymbol{\beta}}_{0w} + (\bar{G}_i - \hat{\gamma}_{i0w}\bar{g}_{iw})\,\hat{\delta}_{0w}, \tag{3.3}$$

where $\hat{\gamma}_{i0w} = \hat{\sigma}_{v0}^2 \big/ \left(\hat{\sigma}_{v0}^2 + \delta_i^2\hat{\sigma}_{e0}^2\right), \bar{g}_{iw} = \sum_{j=1}^{n_i} \tilde{w}_{j|i}\, g_{ij}$ and $\left(\hat{\boldsymbol{\beta}}_{0w}, \hat{\delta}_{0w}\right)$ are obtained by suitably modifying (2.5).

The MSE estimators of $\hat{\mu}_{i(a)}^{H}$ and $\hat{\mu}_{i(a)}^{\mathrm{YR}}$ under the augmented model (1.4) are obtained by suitably modifying (2.8) and (2.9) respectively. Note that we only need to apply existing formulae to the augmented sample model (1.4) to get the EBLUP and the pseudo-EBLUP estimators and associated MSE estimators. New software development is not needed.

Our main interest is to study the performance of the estimators of $\bar{Y}_i$ based on the sample augmented model under informative sampling. Since the estimators $\hat{\bar{Y}}_{i(a)}^{H}$ and $\hat{\mu}_{i(a)}^{\mathrm{YR}}$ are obtained under the augmented model (1.4), they are likely to perform well for the following reasons: (a) If the augmented model holds for the sample, then it also holds for the population, and the non-sampled values y_{ij} can be predicted by fitting the augmented model to the sample; (b) If the augmenting variable g_{ij} explains y_{ij} after conditioning on $\mathbf{x}_{ij}$, then σ^2_{e0} and σ^2_{v0} may be smaller than the corresponding σ^2_e and σ^2_v for the original population model, thus leading to better predictors of the non-sampled y_{ij}. Pfeffermann and Sverchkov (2003) demonstrated, under a different model setup, that the inclusion of sample selection probabilities in the model "can reduce the RMSE quite substantially".

4 Simulation study

4.1 Implementation

A design-model (pm) approach was used for the simulation study by generating data for the $N = \sum_{i=1}^{M} N_i$ population units according to a specified model, and then selecting a sample according to a specified design. The process of generating population data and then selecting a sample is repeated R times. We next describe the steps to implement the process. The population data, y_{ij}, for $M = 99$ areas and $N_i = 100$ units within each area i were generated from the simple nested error linear regression model

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \nu_i + e_{ij};\ i = 1, \ldots, 99;\ j = 1, \ldots, 100, \tag{4.1}$$

where $\beta_0 = 1, \beta_1 = 1, v_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2_v = 0.5\right)$ and independent of $e_{ij} \overset{\text{iid}}{\sim} N\left(0, \sigma^2_e = 2\right)$. The population x_{ij} – values were generated from a gamma distribution with mean 10 and variance 50, and held fixed over the simulation of population y_{ij} – values from (4.1).

We considered different sample sizes within areas by fixing $n_i = 5$ for the first 33 areas, $n_i = 7$ for the next 33 areas and $n_i = 9$ for the last 33 areas. This was done to study the effect of unequal sample sizes on the choice of the augmenting variable $g_{ij} = g\left(p_{j|i}\right)$. Samples of specified sizes, n_i, were selected within the areas with probabilities proportional to specified sizes, b_{ij}, using the Rao-Sampford (Rao 1965 and Sampford 1967) method of sampling with unequal probabilities and without replacement. The latter method ensures that the inclusion probabilities $\pi_{j|i}$ are proportional to the sizes b_{ij}, i.e., $\pi_{j|i} = n_i b_{ij} / B_i = n_i p_{j|i}$, $j = 1, \ldots, N_i$, where B_i is the total of the b_{ij} in area i.

We considered two different choices of the sizes b_{ij} in the simulation study. The first choice uses

$$\begin{aligned} b_{ij} &= \exp\left[\left\{-\left(y_{ij} - \beta_0 - \beta_1 x_{ij}\right) / \sigma_e + \delta_{ij} / 5\right\} / 3\right] \\ &= \exp\left[\left\{-\left(v_i + e_{ij}\right) / \sigma_e + \delta_{ij} / 5\right\} / 3\right], \end{aligned} \tag{4.2}$$

where $\delta_{ij} \overset{\text{iid}}{\sim} N(0,1)$. The size measures (4.2) are equivalent to those used by Pfeffermann and Sverchkov (2007) in their simulation study and satisfy the relationship (1.2) on the weights $w_{j|i} = \pi_{j|i}^{-1}$. Following PS, the area effects v_i and the unit errors e_{ij} were truncated to $\pm 2.5\sigma_v$ and $\pm 2.5\sigma_e$ to avoid extreme selection probabilities.

The second choice of size measures, following Asparouhov (2006), involves two different types of size measures: invariant (I) and non-invariant (NI). For the invariant case, b_{ij} is independent of v_i given $\mathbf{x}_{ij}$; otherwise, it is called non-invariant. Invariant size measures are given by

$$b_{ij} = \left[1 + \exp\left\{-\tau\left(\frac{1}{\alpha} e_{ij} + \sqrt{1 - \frac{1}{\alpha^2}}\, e_{ij}^{*}\right)\right\}\right]^{-1}. \tag{4.3}$$

Non-invariant size measures are taken as

$$b_{ij} = \left[1 + \exp\left\{-\tau\left(\frac{1}{\alpha}\left(v_i + e_{ij}\right) + \sqrt{1 - \frac{1}{\alpha^2}}\left(v_i^{*} + e_{ij}^{*}\right)\right)\right\}\right]^{-1}. \tag{4.4}$$

The coefficient τ in (4.3) and (4.4), chosen as 0.5, ensures that the variation of the weights $w_{j|i}$ would not be too large within a simulation run. The random pair $\left(v_i^{*}, e_{ij}^{*}\right)$ was generated independently of $\left(v_i, e_{ij}\right)$ from the same distributions as v_i and e_{ij} to ensure that the weight variation would be comparable between various levels of α. If some of the $\pi_{j|i}$ exceeded one, they were set to one, and the probabilities were recomputed for the remaining units. The $\alpha-$values in (4.3) and (4.4), chosen as 1, 2, 3 or ∞, control the level of informativeness. Increasing α decreases informativeness, with $\alpha = \infty$ corresponding to non-informative sampling. Various dependencies in the simulations were introduced as follows, in order to increase the precision of comparisons between different estimators: All the four error components $\left(v_i, e_{ij}, v_i^{*}, e_{ij}^{*}\right)$ were first generated. Population $y-$values, as well as invariant and non-invariant probabilities of selection, were then generated from those errors. For a given generated population, eight samples were selected: an invariant sample and a non-invariant sample for each value of α considered.

It may be noted that the weights $w_{j|i}$ obtained from the size measures (4.3) and (4.4) may not satisfy condition (1.2) of PS. We nevertheless fitted (1.2) to those weights to compute $\hat{b}$ needed in the bias-adjusted estimator $\hat{\bar{Y}}_i^{\text{PS}}$.

Using the design-model (pm) approach, $R = 1{,}000$ samples were generated under the size measures (4.2) and the size measures (4.3) and (4.4). From each simulated sample $r\,(r = 1, \ldots, R)$, the estimates $\hat{\bar{Y}}_i^{H(r)}, \hat{\bar{Y}}_{i(a)}^{H(r)}$ and $\hat{\bar{Y}}_i^{\text{PS}(r)}$ were computed for each small area i; for the YR method only $\hat{\mu}_i^{\text{YR}(r)}$ and $\hat{\mu}_{i(a)}^{\text{YR}(r)}$ were computed. Also, the MSE estimates, $\text{mse}\left(\hat{\mu}_i^{H}\right)^{(r)}, \text{mse}\left(\hat{\mu}_{i(a)}^{H}\right)^{(r)}, \text{mse}\left(\hat{\mu}_i^{\text{YR}}\right)^{(r)}$ and $\text{mse}\left(\hat{\mu}_{i(a)}^{\text{YR}}\right)^{(r)}$, associated with $\hat{\mu}_i^{H}, \hat{\mu}_{i(a)}^{H}, \hat{\mu}_i^{\text{YR}}$ and $\hat{\mu}_{i(a)}^{\text{YR}}$, were computed. As noted earlier, we did not include the bootstrap MSE estimator of $\hat{\bar{Y}}_i^{\text{PS}}$, proposed by Pfeffermann and Sverchkov (2007), in the simulation study. Also, for simplicity, we did not include the MSE estimators of $\hat{\bar{Y}}_i^{H}$ and $\hat{\bar{Y}}_{i(a)}^{H}$ because the latter estimators performed similarly to $\hat{\mu}_i^{H}$ and $\hat{\mu}_{i(a)}^{H}$ in terms of MSE.

We considered the following performance measures for a given estimator, say of the small area mean $\bar{Y}_i$. Average absolute bias $\left(\overline{\mathrm{AB}}\right)$ is measured by

$$\overline{\mathrm{AB}} = \frac{1}{M}\sum_{i=1}^{M}\mathrm{AB}_i$$

with

$$\mathrm{AB}_i = \left|\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\bar{Y}}_i^{(r)} - \bar{Y}_i^{(r)}\right)\right|$$

where $\hat{\bar{Y}}_i^{(r)}$ and $\bar{Y}_i^{(r)}$ are the values of $\hat{\bar{Y}}_i$ and $\bar{Y}_i$ for the r^{th} simulated sample and population. Efficiency of an estimator $\hat{\bar{Y}}_i$ is measured by the average root MSE

$$\overline{\mathrm{RMSE}} = \frac{1}{M}\sum_{i=1}^{M}\sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\bar{Y}}_i^{(r)} - \bar{Y}_i^{(r)}\right)^2}.$$

Turning to the performance of MSE estimators $\mathrm{mse}\left(\hat{\mu}_i^H\right), \mathrm{mse}\left(\hat{\mu}_{i(a)}^H\right), \mathrm{mse}\left(\hat{\mu}_i^{\mathrm{YR}}\right)$ and $\mathrm{mse}\left(\hat{\mu}_{i(a)}^{\mathrm{YR}}\right)$ in estimating MSEs, we first calculated reliable measures of MSEs by increasing $R = 1{,}000$ to $T = 10{,}000$ simulated samples. The MSE of an estimator $\hat{\mu}_i$ is then calculated as

$$\mathrm{MSE}\left(\hat{\mu}_i\right) = \frac{1}{T}\sum_{t=1}^{T}\left(\hat{\mu}_i^{(t)} - \bar{Y}_i^{(t)}\right)^2,$$

where $\hat{\mu}_i^{(t)}$ and $\bar{Y}_i^{(t)}$ denote the values of $\hat{\mu}_i$ and $\bar{Y}_i$ for the t^{th} simulated sample and population. For MSE estimation, we retained the original R simulated samples and calculated the expected values $E\left[\mathrm{mse}\left(\hat{\mu}_i\right)\right] = R^{-1}\sum_{r=1}^{R}\mathrm{mse}\left(\hat{\mu}_i\right)^{(r)}$, where $\mathrm{mse}\left(\hat{\mu}_i\right)^{(r)}$ denotes the value of the MSE estimate for the r^{th} simulated sample. The average absolute relative bias $\left(\overline{\mathrm{ARB}}\right)$ of a MSE estimator $\mathrm{mse}\left(\hat{\mu}_i\right)$ is then calculated as

$$\overline{\mathrm{ARB}}\left[\mathrm{mse}\left(\hat{\mu}_i\right)\right] = M^{-1}\sum_{i=1}^{M}\left|\frac{E\left[\mathrm{mse}\left(\hat{\mu}_i\right)\right]}{\mathrm{MSE}\left(\hat{\mu}_i\right)} - 1\right|.$$

4.2 Results under the Pfeffermann and Sverchkov size measures

Table 4.1 reports the simulation results on the average absolute bias $\left(\overline{\mathrm{AB}}\right)$ and the average root mean square error $\left(\overline{\mathrm{RMSE}}\right)$ of the estimators $\hat{\bar{Y}}_i^H, \hat{\bar{Y}}_{i(a)}^H, \hat{\mu}_i^{\mathrm{YR}}, \hat{\mu}_{i(a)}^{\mathrm{YR}}$ and $\hat{\bar{Y}}_i^{\mathrm{PS}}$ under the PS size measures (4.2). The average absolute RB $\left(\overline{\mathrm{ARB}}\right)$ of the MSE estimators $\mathrm{mse}\left(\hat{\mu}_i^H\right), \mathrm{mse}\left(\hat{\mu}_{i(a)}^H\right), \mathrm{mse}\left(\hat{\mu}_i^{\mathrm{YR}}\right)$ and $\mathrm{mse}\left(\hat{\mu}_{i(a)}^{\mathrm{YR}}\right)$ are also reported. Four different choices of the augmenting variable g_{ij} were studied: $p_{j|i}, w_{j|i}, n_i w_{j|i} = p_{j|i}^{-1}$ and $\log p_{j|i}$. Bootstrap estimator of $\mathrm{MSE}\left(\hat{\bar{Y}}_i^{\mathrm{PS}}\right)$, proposed by Pfeffermann and Sverchkov (2007), is not included in our study because the bootstrap simulation is very computer intensive.

Table 4.1 shows that the $\overline{\text{AB}}$ of the EBLUP estimator $\hat{\hat{Y}}_i^H$ is large $(= 0.456)$ relative to the corresponding augmented model EBLUP, $\hat{\hat{Y}}_{i(a)}^H$, for the four choices of g_{ij}. Also, the choice $g_{ij} = w_{j|i}$ leads to larger $\overline{\text{AB}}$ compared to the other three choices (0.131 compared to 0.042 or less). The customary pseudo-EBLUP, $\hat{\mu}_i^{\text{YR}}$, surprisingly performed well $\left(\overline{\text{AB}} = 0.044\right)$ even though it was obtained under the assumption of noninformative sampling. This good performance is perhaps due to the use of weights in $\hat{\mu}_i^{\text{YR}}$. Augmented pseudo-EBLUP, $\hat{\mu}_{i(a)}^{\text{YR}}$, leads to further reduction in $\overline{\text{AB}}$. The PS estimator, $\hat{\hat{Y}}_i^{\text{PS}}$, performs well relative to $\hat{\hat{Y}}_{i(a)}^H$: $\overline{\text{AB}} = 0.033$.

Turning to $\overline{\text{RMSE}}$, Table 4.1 shows that $\hat{\hat{Y}}_i^H$ has the largest value $(= 0.617)$ due to large $\overline{\text{AB}}$, followed by $\hat{\mu}_i^{\text{YR}}$ and $\hat{\hat{Y}}_i^{\text{PS}}$ with values 0.442 and 0.416 respectively. On the other hand, the augmented model estimators performed significantly better relative to $\hat{\hat{Y}}_i^{\text{PS}}$ and $\hat{\mu}_i^{\text{YR}}$. For example, the choice $g_{ij} = p_{j|i}$ gives $\overline{\text{RMSE}} = 0.151$. Among the four choices of g_{ij}, the choice $w_{j|i}$ gives the largest $\overline{\text{RMSE}}\,(= 0.242)$. We also calculated $\overline{\text{AB}}$ and $\overline{\text{RMSE}}$ of the approximate EBLUP estimators $\hat{\mu}_i^H$ and $\hat{\mu}_{i(a)}^H$. We found that the values are practically the same as the corresponding values for $\hat{\hat{Y}}_i^H$ and $\hat{\hat{Y}}_{i(a)}^H$.

Finally, with respect to MSE estimation, $\text{mse}\left(\hat{\mu}_i^H\right)$ exhibits largest $\overline{\text{ARB}}$: 53.1% compared to 3.8% for $\hat{\mu}_i^{\text{YR}}$, although $\overline{\text{RMSE}}$ for $\hat{\mu}_i^{\text{YR}}$ is larger compared to $\hat{\mu}_{i(a)}^H$ based on $p_{j|i}$ or $n_i w_{j|i}$. The MSE estimators $\text{mse}\left(\hat{\mu}_{i(a)}^H\right)$ and $\text{mse}\left(\hat{\mu}_{i(a)}^{\text{YR}}\right)$ lead to small $\overline{\text{ARB}}\,(<\ 7\%)$ except for the choice $w_{j|i}$ which leads to $\overline{\text{ARB}} = 62.6\%$ for $\hat{\mu}_{i(a)}^H$ and $\overline{\text{ARB}} = 39.6\%$ for $\hat{\mu}_{i(a)}^{\text{YR}}$.

Table 4.1
Average absolute bias $\left(\overline{\text{AB}}\right)$, average RMSE $\left(\overline{\text{RMSE}}\right)$ of the estimators and percent average absolute RB $\left(\overline{\text{ARB}}\right)$ of the MSE estimators: PS size measures

Performance measure	EBLUP					pseudo-EBLUP					PS
	$\hat{\hat{Y}}_i^H$	$\hat{\hat{Y}}_{i(a)}^H$				$\hat{\mu}_i^{\text{YR}}$	$\hat{\mu}_{i(a)}^{\text{YR}}$				$\hat{\hat{Y}}_i^{\text{PS}}$
		$p_{j\|i}$	$n_i w_{j\|i}$	$w_{j\|i}$	$\log p_{j\|i}$		$p_{j\|i}$	$n_i w_{j\|i}$	$w_{j\|i}$	$\log p_{j\|i}$	
$\overline{\text{AB}}$	0.456	0.042	0.004	0.131	0.003	0.044	0.007	0.004	0.044	0.003	0.033
$\overline{\text{RMSE}}$	0.617	0.151	0.147	0.242	0.101	0.442	0.157	0.156	0.207	0.106	0.416
$\%\overline{\text{ARB}}(\text{mse})$	53.1	3.7	6.7	62.6	6.9	3.8	4.1	5.2	39.6	6.7	

4.3 Selection of the augmenting variable

In this section we illustrate the selection of the augmenting variable by generating data for the N population units from model (4.1) and then selecting a sample from the population data according to the Rao-Sampford method using size measures (4.2). Letting $u_{ij} = v_i + e_{ij}$, we fitted the model $y_{ij} = \beta_0 + \beta_1 x_{ij} + u_{ij}$ to the sample data by ordinary least squares (OLS) and obtained the residuals $\tilde{u}_{ij} = y_{ij} - \tilde{\beta}_0 - \tilde{\beta}_1 x_{ij}$, where $\tilde{\beta}_0$ and $\tilde{\beta}_1$ are the OLS estimators of β_0 and β_1 respectively.

Figure 4.1 gives residual plots of $\left(\tilde{u}_{ij}, p_{j|i}\right), \left(\tilde{u}_{ij}, \log p_{j|i}\right), \left(\tilde{u}_{ij}, n_i w_{j|i}\right)$ and $\left(\tilde{u}_{ij}, w_{j|i}\right)$. All four plots clearly indicate informative sampling. Linear relationships between u_{ij} and the two choices $p_{j|i}$ and

$\log p_{j|i}$ suggest that either of them should work well. The choice $w_{j|i}$ indicates some non-linearity and wider scatter in the residual plot, and this choice led to the largest $\overline{\text{RMSE}}$ among the four choices, as shown in Table 4.1. The choice $n_i w_{j|i}$ also indicates some non-linearity but less scatter in the residual plot.

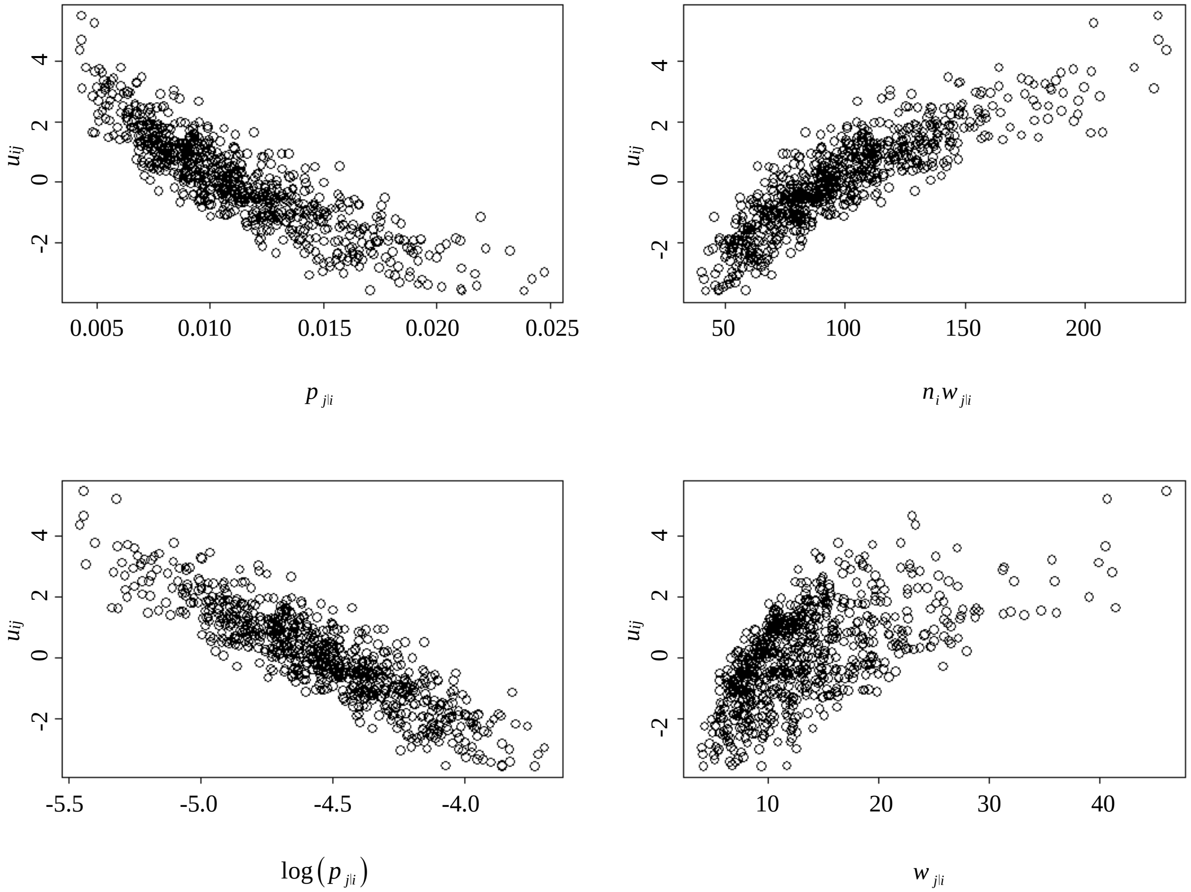


Figure 4.1 Residual plots for a simulated example: PS size measures.

We have also fitted the augmented model (1.4) with $g(p_{j|i}) = p_{j|i}$ and calculated the OLS residuals $\tilde{u}_{0ij} = y_{ij} - \tilde{\beta}_{00} - \tilde{\beta}_{01}x_{ij} - \tilde{\delta}_0 p_{j|i}$. All the residuals $\tilde{u}_{0ij}$ are less than 2.0 in absolute value, suggesting adequacy of the augmented model.

4.4 Results under Asparouhov size measures

Table 4.2 reports the simulation results on $\overline{\text{AB}}$ under the Asparouhov size measures (4.3) and (4.4). It shows, as in Table 4.1 for the PS size measures, that $\overline{\text{AB}}$ of the EBLUP is large (0.437 for the invariant size measures (I) and 0.440 for non-invariant size measures (NI)) when the augmenting variable, g_{ij}, is not included in the model and sampling is very informative $(\alpha = 1)$. Also, $\overline{\text{AB}}$ decreases as α increases. On the other hand, under the same model $\overline{\text{AB}}$ associated with pseudo-EBLUP is much lower: 0.048 for I and 0.047 for NI when $\alpha = 1$, and $\overline{\text{AB}}$ decreases as α increases. The PS estimator under the same model also exhibits lower $\overline{\text{AB}}$ (about 0.01) regardless of the choice of the value of α. Inclusion of $p_{j|i}$ or $n_i w_{j|i}$ or $\log p_{j|i}$ as augmenting variable in the model also leads to small $\overline{\text{AB}}$ for the EBLUP (0.02 or less) regardless of the value of α. On the other hand, the choice $w_{j|i}$ as the augmenting variable leads to larger $\overline{\text{AB}}$ (0.14 for $\alpha = 1$ and 2), except for non-informative sampling $(\alpha = \infty)$. This poor performance of the choice $w_{j|i}$ is probably due to the fact that $w_{j|i} = \left(n_i p_{j|i}\right)^{-1}$ depends on n_i when the area sample sizes, n_i, are not equal, unlike the other choices of g_{ij}. Pseudo-EBLUP performed similarly to EBLUP under the augmented model in terms of $\overline{\text{AB}}$.

Table 4.2
Average absolute bias $\left(\overline{\text{AB}}\right)$ of the estimators under Asparouhov size measures: invariant (I) and noninvariant (NI)

α	Size measure	EBLUP					pseudo-EBLUP					PS
		$\hat{\bar{Y}}_i^H$	$\hat{\bar{Y}}_{i(a)}^H$				$\hat{\mu}_i^{\text{YR}}$	$\hat{\mu}_{i(a)}^{\text{YR}}$				$\hat{\bar{Y}}_i^{\text{PS}}$
			$p_{j\|i}$	$n_i w_{j\|i}$	$w_{j\|i}$	$\log p_{j\|i}$		$p_{j\|i}$	$n_i w_{j\|i}$	$w_{j\|i}$	$\log p_{j\|i}$	
1	I	0.437	0.001	0.005	0.140	0.022	0.048	0.001	0.006	0.057	0.005	0.012
	NI	0.440	0.007	0.007	0.145	0.021	0.047	0.003	0.007	0.064	0.005	0.013
2	I	0.217	0.009	0.010	0.137	0.014	0.024	0.010	0.010	0.098	0.010	0.012
	NI	0.217	0.011	0.009	0.136	0.011	0.024	0.009	0.010	0.098	0.010	0.012
3	I	0.145	0.010	0.010	0.101	0.011	0.017	0.010	0.010	0.075	0.010	0.011
	NI	0.144	0.011	0.011	0.099	0.012	0.016	0.010	0.011	0.074	0.011	0.011
∞	I	0.011	0.011	0.011	0.011	0.011	0.012	0.011	0.011	0.012	0.011	0.011
	NI	0.010	0.010	0.010	0.010	0.010	0.010	0.010	0.010	0.010	0.010	0.010

Table 4.3 reports the simulation results on the average root mean squared error $\left(\overline{\text{RMSE}}\right)$ using the Asparouhov size measures (4.3) and (4.4). It shows that the EBLUP, based on model (1.4) without the augmenting variable g_{ij}, has the largest $\overline{\text{RMSE}}$ (0.596 for I and 0.619 for NI) when the sampling is very informative $(\alpha = 1)$. The $\overline{\text{RMSE}}$ gradually decreases to around 0.42 as the sampling becomes non-informative $(\alpha = \infty)$. On the other hand, $\overline{\text{RMSE}}$ of both the pseudo-EBLUP (without the g_{ij} – term in the model) and PS do not depend on α, and lead to significant reduction: $\overline{\text{RMSE}}$ of the pseudo-EBLUP is around 0.44 and $\overline{\text{RMSE}}$ of PS is slightly smaller, around 0.42. Increase in $\overline{\text{RMSE}}$ of the pseudo-EBLUP and PS over the EBLUP under non-informative sampling $(\alpha = \infty)$ is also small. On the other hand, EBLUP and pseudo-EBLUP under the augmented model lead to large reduction in MSE when the sampling is very informative $(\alpha = 1)$, particularly for the choices $p_{j|i}$ and $\log p_{j|i}$: $\overline{\text{RMSE}}$ less than 0.15. The choice of $w_{j|i}$ leads to larger $\overline{\text{RMSE}}$ (around 0.29) when $\alpha = 1$ but it is still much smaller than the $\overline{\text{RMSE}}$ for the pseudo-EBLUP without the g_{ij} – term and the PS. As α increases, $\overline{\text{RMSE}}$ is roughly the same for EBLUP (under the augmented model), pseudo-EBLUP and PS.

Table 4.3
Average root mean squared error $\left(\overline{\text{RMSE}}\right)$ of the estimators under Asparouhov size measures: invariant (I) and noninvariant (NI)

α	Size measure	EBLUP					pseudo-EBLUP					PS
		$\hat{\bar{Y}}_i^H$	$\hat{\bar{Y}}_{i(\mathrm{a})}^H$				$\hat{\mu}_i^{\mathrm{YR}}$	$\hat{\mu}_{i(\mathrm{a})}^{\mathrm{YR}}$				$\hat{\bar{Y}}_i^{\mathrm{PS}}$
			$p_{j\|i}$	$n_i w_{j\|i}$	$w_{j\|i}$	$\log p_{j\|i}$		$p_{j\|i}$	$n_i w_{j\|i}$	$w_{j\|i}$	$\log p_{j\|i}$	
1	I	0.596	0.039	0.203	0.281	0.108	0.454	0.040	0.223	0.258	0.112	0.406
	NI	0.619	0.110	0.205	0.295	0.135	0.457	0.092	0.235	0.273	0.136	0.435
2	I	0.468	0.377	0.385	0.418	0.379	0.436	0.391	0.398	0.415	0.392	0.416
	NI	0.474	0.375	0.378	0.414	0.374	0.438	0.392	0.396	0.413	0.391	0.423
3	I	0.439	0.400	0.403	0.420	0.401	0.432	0.414	0.417	0.425	0.415	0.415
	NI	0.443	0.400	0.401	0.418	0.399	0.435	0.416	0.416	0.425	0.415	0.420
∞	I	0.417	0.418	0.418	0.418	0.418	0.431	0.431	0.431	0.432	0.431	0.418
	NI	0.418	0.418	0.418	0.419	0.418	0.432	0.432	0.432	0.433	0.432	0.418

Table 4.4 reports the simulation result on the average absolute relative bias $\left(\overline{\text{ARB}}\right)$ of MSE estimators under the Asparouhov size measures (4.3) and (4.4). It shows that $\overline{\text{ARB}}$ of the MSE estimator of the EBLUP, based on the model without the augmenting variable g_{ij}, is very large when the sampling is very informative $(\alpha = 1)$: 52.8% for I and 59.1% for NI. $\overline{\text{ARB}}$ gradually decreases to around 5% under non-informative sampling $(\alpha = \infty)$. The use of $\log p_{j|i}$ as an augmenting variable leads to large reduction in $\overline{\text{ARB}}$ $(< 9\%)$ and the choices $p_{j|i}$ and $n_i w_{j|i}$ also perform well in terms of $\overline{\text{ARB}}$ except for the case of NI and $\alpha = 1$ which leads to 18.5% and 12.9% respectively. Again, $w_{j|i}$ is not a good choice because it leads to $\overline{\text{ARB}}$ as large as 40% when $\alpha = 1.$ The MSE estimator associated with the pseudo-EBLUP (without g_{ij}) also performs well, except for NI and $\alpha = 1,$ leading to $\overline{\text{ARB}}$ of 19.5%. Use of $\log p_{j|i}$ as auxiliary variable leads to $\overline{\text{ARB}}$ less than 8% for the MSE estimator associated with the pseudo-EBLUP. We have not included the PS bootstrap MSE estimator in our study.

Overall, our simulation study indicates that the use of augmented models under informative sampling leads to EBLUPs that perform well in terms of $\overline{\text{AB}}$ and $\overline{\text{RMSE}}$ of the estimators, and $\overline{\text{ARB}}$ of MSE estimators, provided that the augmenting variable is chosen properly. The bias-adjusted estimators of PS also perform well, even though they led to larger $\overline{\text{RMSE}}$ under the PS size measures (4.2). Pseudo-EBLUP estimators (without the augmenting variable) also perform well and further improvement may be achieved under augmented models.

Table 4.4
Average relative bias (%) of MSE estimators under Asparouhov size measures: invariant (I) and noninvariant (NI)

α	Size measure	EBLUP					pseudo-EBLUP				
		$\hat{\bar{Y}}_i^H$	$\hat{\bar{Y}}_{i(\mathrm{a})}^H$				$\hat{\mu}_i^{\mathrm{YR}}$	$\hat{\mu}_{i(\mathrm{a})}^{\mathrm{YR}}$			
			$p_{j\|i}$	$n_i w_{j\|i}$	$w_{j\|i}$	$\log p_{j\|i}$		$p_{j\|i}$	$n_i w_{j\|i}$	$w_{j\|i}$	$\log p_{j\|i}$
1	I	52.8	6.5	4.8	39.8	3.3	11.7	6.6	7.8	19.2	6.2
	NI	59.1	18.5	12.9	39.4	7.8	19.5	26.0	10.2	16.6	6.0
2	I	19.4	6.0	5.5	10.7	5.9	3.9	6.3	6.0	7.3	6.4
	NI	22.6	8.8	8.0	11.3	8.6	4.2	6.7	6.0	7.4	6.7
3	I	7.1	5.5	5.5	5.3	5.5	4.4	6.0	6.3	7.2	6.3
	NI	8.9	7.3	7.0	5.9	7.2	4.0	7.1	7.0	7.3	7.2
∞	I	5.1	5.1	5.0	5.0	5.1	5.1	5.2	5.3	5.3	5.2
	NI	5.0	4.9	4.9	4.9	4.9	4.9	5.0	5.1	5.1	5.0

5 Concluding remarks

In this paper, we studied model-based small area estimation for different levels of design informativeness under a nested error linear regression model for the population units. Estimators considered were the EBLUP, the pseudo-EBLUP (You and Rao 2002) and an estimator given by Pfeffermann and Sverchkov (2007). The EBLUP and the pseudo-EBLUP were computed under two scenarios: (i) Ignore informative sampling and assume that the population model holds for the sample; (ii) Take account of informative sampling by using a suitable function of the unit selection probability $p_{j|i}$ as an additional auxiliary variable in the sample model.

Results from a simulation study showed that design informativeness can have a big impact on the bias and MSE of the EBLUP that ignores informative sampling (scenario (i)). Results under scenario (ii) showed that the EBLUP, based on the augmented model, performs extremely well in terms of bias and MSE, provided that the augmenting variable is chosen properly. The bias-adjusted estimator of Pfeffermann and Sverchkov (2007) also performed well under informative sampling in terms of bias but its MSE is significantly larger than the corresponding MSE of the EBLUP and the pseudo-EBLUP based on the augmented model. Pseudo-EBLUP under scenario (i) performed significantly better than the corresponding EBLUP. It can be significantly improved by using the augmented model, similar to the case of EBLUP.

An advantage of the augmented model approach is that no new theory is required for estimation and MSE estimation. However, the area mean $\bar{G}_i$ of the augmenting variable g_{ij} is required, unlike in the approach of Pfeffermann and Sverchkov (2007). For some choices of $g_{ij}, \bar{G}_i$ is readily known; for example $g_{ij} = p_{j|i}$ gives $\bar{G}_i = 1/N_i$ and $g_{ij} = n_i w_{j|i}$ gives $\bar{G}_i = n_i \bar{W}_i$ and $\bar{W}_i$ is often known for some surveys. We have also given a method of choosing the augmenting variable g_{ij}.

In this paper, we focused on the special case where all the areas are sampled. Extension of the augmented model approach to handle non-sampled areas requires the knowledge of the area means $\bar{G}_i$, as well as the area selection probabilities, p_i, for the non-sampled areas. This extension is currently under study.

Acknowledgements

We are thankful to the Associate Editor, the referees and M. Sverchkov for many constructive comments and suggestions.

References

Asparouhov, T. (2006). General multi-level modelling with sampling weights. *Communication in Statistics, Theory and Methods,* 439-460.

Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An error component model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.

Henderson, C.R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics,* 31, 423-447.

Pfeffermann, D., and Sverchkov, M. (2003). Small area estimation under informative sampling. *Proceedings of the Survey Research Methods Section,* American Statistical Association, 3284-3295.

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, 480, 1427-1439.

Prasad, N.G.N., and Rao, J.N.K. (1999). On robust small area estimation using a simple random effects model. *Survey Methodology,* 25, 1, 67-72.

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

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

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

Skinner, C.J. (1994). Sampling models and weights. *Proceedings of the Survey Research Methods Section,* American Statistical Association, 133-142.

Stefan, M. (2005). Contributions à l’estimation pour petits domaines. Ph.D. thesis, Université Libre de Bruxelles.

Torabi, M., and Rao, J.N.K. (2010). Mean squared error estimators of small area means using survey weights. *The Canadian Journal of Statistics,* 38, 4, 595-608.

You, Y., and Rao, J.N.K. (2002). A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *The Canadian Journal of Statistics,* 30, 3, 431-439.

Vol. 41, No. 2, pp. 349-376
Statistics Canada, Catalogue No. 12-001-X

Combining link-tracing sampling and cluster sampling to estimate the size of a hidden population in presence of heterogeneous link-probabilities

Martín H. Félix-Medina, Pedro E. Monjardin and Aida N. Aceves-Castro[1]

Abstract

Félix-Medina and Thompson (2004) proposed a variant of link-tracing sampling to sample hidden and/or hard-to-detect human populations such as drug users and sex workers. In their variant, an initial sample of venues is selected and the people found in the sampled venues are asked to name other members of the population to be included in the sample. Those authors derived maximum likelihood estimators of the population size under the assumption that the probability that a person is named by another in a sampled venue (link-probability) does not depend on the named person (homogeneity assumption). In this work we extend their research to the case of heterogeneous link-probabilities and derive unconditional and conditional maximum likelihood estimators of the population size. We also propose profile likelihood and bootstrap confidence intervals for the size of the population. The results of simulations studies carried out by us show that in presence of heterogeneous link-probabilities the proposed estimators perform reasonably well provided that relatively large sampling fractions, say larger than 0.5, be used, whereas the estimators derived under the homogeneity assumption perform badly. The outcomes also show that the proposed confidence intervals are not very robust to deviations from the assumed models.

Key Words: Bootstrap; Capture-recapture; Chain referral sampling; Maximum likelihood estimator; Profile likelihood confidence interval; Snowball sampling.

1 Introduction

Conventional sampling methods are not appropriate for sampling hidden or hard-to-reach human populations, such as drug users, sexual-workers and homeless people, because of the lack of suitable sampling frames. For this reason, several specific sampling methods for this type of population have been proposed. See Magnani, Sabin, Saidel and Heckathorn (2005) and Kalton (2009) for reviews of some of them. According to Heckathorn (2002) two types of sampling methods for hidden populations are the most commonly used in actual studies. One is location sampling, also known as time-and-space sampling, aggregation point sampling or intercept point sampling. The other is snowball sampling, also known as link-tracing sampling (LTS) or chain referral sampling.

In location sampling a frame of primary units is constructed. The primary units are combinations of places and time segments where the elements of the population tend to gather. The frame is not assumed to cover the whole population. A probability sample of primary units is selected and from each sampled unit a sort of systematic sample of elements is drawn. Although design-based estimators of different parameters can be constructed, the main drawback of location sampling is that inferences are valid only for the part of the population covered by the frame. For reviews of this method see MacKellar, Valleroy, Karon, Lemp and Janssen (1996), Munhib, Lin, Stueve, Miller, Ford, Johnson and Smith (2001), McKenzie and Mistianen (2009), Semaan (2010) and Karon and Wejnert (2012). Location sampling has been used in the Young Men’s Survey to estimate HIV seroprevalence in young men who have sex with

[1] Martín H. Félix-Medina, Pedro E. Monjardin and Aida N. Aceves-Castro, Facultad de Ciencias F sico-Matemáticas, Universidad Autónoma de Sinaloa, Ciudad Universitaria, Culiacán Sinaloa, México. E-mail: mhfelix@uas.edu.mx.

men. (See McKellar et al. 1996.) In this study the primary units were venues attended by young men such as dance clubs, bars and street locations.

In LTS an initial sample of members of the population is selected and the sample size is increased by asking the sampled people to name or to refer other members of the population to be included in the sample. The named people who are not in the initial sample might be asked to refer other persons, and the process might continue in this way until a specified stopping rule is satisfied. For reviews of several variants of LTS see Spreen (1992), Thompson and Frank (2000) and Johnston and Sabin (2010). LTS was used in the Colorado Springs study on heterosexual transmission of HIV/AIDS. (See Potterat, Woodhouse, Rothenberg, Muth, Darrow, Muth and Reynolds 1993; Rothemberg, Woodhouse, Potterat, Muth, Darrow and Klovdahl 1995 and Potterat, Woodhouse, Muth, Rothenberg, Darrow, Klovdahl and Muth 2004.) In this research an initial non probabilistic sample of people presumably at high risk of acquiring and transmitting HIV was obtained and they were asked for a complete enumeration of their personal contacts who were also included in the sample.

Frank and Snijders (1994) proposed a variant of LTS that allows the sampler to estimate the population size. In their variant they assume an initial Bernoulli sample, that is, that every element of the population has the same probability of being included in the sample and that the inclusions are independent. In addition, they assume that the probability that person i in the initial sample refers person j in the population, which we will call link-probability, is a constant, and that the referrals are independent. We will name the first of the additional premises the assumption of homogeneity of the link-probabilities. Based on these hypotheses these authors derive several estimators of the population size. They indicate that their method yielded reasonable estimates of the number of heroin users in Groningen. However, Dávid and Snijders (2002) reported an underestimate of the number of homeless in Budapest using this method. They indicate that the underestimation might be caused by deviations from the assumption of an initial Bernoulli sample.

The problem of satisfying in actual applications of LTS the assumption of an initial Bernoulli sample of members of the population motivated Félix-Medina and Thompson (2004) to develop a variant of LTS in which the initial sample is selected by a probabilistic design. To do this they assume, as in location sampling, that the sampler can construct a sampling frame of sites or venues where the members of the population tend to gather, such as bars, parks and blocks. The frame is not assumed to cover the whole population, but only a portion of it. Then, a simple random sample without replacement (SRSWOR) of sites is selected and the members of the population who belong to the sampled sites are identified. Finally, as in ordinary LTS, the people in the initial sample are asked to name other members of the population.

These authors propose maximum likelihood estimators (MLEs) of the population size derived under a probability model that describes the numbers of people found in the sampled sites and a model that regards that the link-probabilities between the elements of the population and the sampled sites are homogeneous, that is, that they depend on the sampled sites, but not on the potentially named people. Later, Félix-Medina and Monjardin (2006) consider this same variant of LTS and propose estimators of the population size derived also under the assumption of homogeneity, but using a Bayesian-assisted approach, that is, the functional forms of the estimators are obtained using the Bayesian approach, but inferences are made under the frequentist approach.

Although the variant of LTS proposed by Félix-Medina and Thompson (2004) has not been used in any actual study, we would expect that estimators of the population size derived under the assumption of homogeneity will present problems of underestimation if this hypothesis is not satisfied as occurs in capture-recapture studies. We think this because these estimators resemble those used in that field.

In this paper, we extend the work by Félix-Medina and Thompson (2004) to the case in which the link-probabilities depend on the named people, that is, we assume heterogeneous link-probabilities. The structure of the paper is as follows. In Section 2 we introduce the LTS variant proposed by Félix-Medina and Thompson (2004). In Section 3 we present a model for the link-probabilities that takes into account their heterogeneity and derive unconditional and conditional MLEs of the population size. In Section 4 we construct profile likelihood and bootstrap confidence intervals for the population size. In Section 5 we present a procedure for determining the size of the initial sample in order to achieve a specified value of the relative error of the estimation. In Section 6 we describe the results of two simulation studies, and finally, in Section 7 we present some conclusions and suggestions for future research.

2 Sampling design and notation

Since in this work we consider the variant of LTS proposed by Félix-Medina and Thompson (2004), we will briefly describe it. Thus, let U be a finite population of an unknown number τ of people. We assume that a portion U_1 of U is covered by a sampling frame of N sites $A_1, \ldots, A_N$, where the members of the population can be found with high probability. We suppose that we have a criterion that allows us to assign a person in U_1 to only one site in the frame. Notice that we are not assuming that a person could not be found in different sites, but that, as in ordinary cluster sampling, we are able to assign him or her to only one site, for instance, the site where he or she spends most of his or her time. Thus, we can consider the sites in the frame as clusters of people. Let M_i denote the number of members of the population that belong to the site $A_i, i = 1, \ldots, N$. From the previous assumption it follows that the number of people in U_1 is $\tau_1 = \sum_1^N M_i$ and the number of people in the portion $U_2 = U - U_1$ of U that is not covered by the frame is $\tau_2 = \tau - \tau_1$.

The sampling design is as follows. A SRSWOR S_A of n sites $A_1, \ldots, A_n$ is selected from the frame and the M_i members of the population who belong to the sampled site A_i are identified, $i = 1, \ldots, n$. Let S_0 be the set of people in the initial sample. Observe that the size of S_0 is $M = \sum_1^n M_i$. The people in each sampled site are asked to name other members of the population. We will say that a person and a site are linked if any of the people who belong to that site names him or her. Finally, let S_1 and S_2 be the sets of people in $U_1 - S_0$ and U_2, respectively, who are linked to some site or sites in S_A.

3 Maximum likelihood estimators of τ_1, τ_2 and τ

3.1 Probability models

To construct MLEs of the τ's we need to specify models for the observed variables. Thus, as in Félix-Medina and Thompson (2004), we will suppose that the numbers $M_1, \ldots, M_N$ of people who belong to the sites $A_1, \ldots, A_N$ are independent Poisson random variables with mean λ_1. Therefore, the joint conditional distribution of $(M_1, \ldots, M_n, \tau_1 - M)$ given that $\sum_1^N M_i = \tau_1$ is multinomial with probability mass function (pmf):

$$f(m_1, \ldots, m_n, \tau_1 - m) = \frac{\tau_1!}{\prod_1^n m_i!(\tau_1 - m)!}\left(\frac{1}{N}\right)^m \left(1 - \frac{n}{N}\right)^{\tau_1 - m}. \tag{3.1}$$

To model the links between the members of the population and the sampled sites we will define the following random variables: $X_{ij}^{(k)} = 1$ if person j in $U_k - A_i$ is linked to site A_i and $X_{ij}^{(k)} = 0$ if $j \in A_i$ or that person is not linked to A_i, $j = 1, \ldots, \tau_k, i = 1, \ldots, n$. We will suppose that given the sample S_A of sites the $X_{ij}^{(k)}$'s are independent Bernoulli random variables with means $p_{ij}^{(k)}$'s, where the link-probability $p_{ij}^{(k)}$ satisfies the following Rasch model:

$$p_{ij}^{(k)} = \Pr\left(X_{ij}^{(k)} = 1 \middle| \beta_j^{(k)}, S_A\right) = \frac{\exp\left(\alpha_i^{(k)} + \beta_j^{(k)}\right)}{1 + \exp\left(\alpha_i^{(k)} + \beta_j^{(k)}\right)}, \; j \in U_k - A_i; \; i = 1, \ldots, n. \tag{3.2}$$

It is worth noting that this model was considered by Coull and Agresti (1999) in the context of capture-recapture sampling. In this model $\alpha_i^{(k)}$ is a fixed (not random) effect that represents the potential that the cluster A_i has of forming links with the people in $U_k - A_i$, and $\beta_j^{(k)}$ is a random effect that represents the propensity of the person $j \in U_k$ to be linked to a cluster. We will suppose that $\beta_j^{(k)}$ is normally distributed with mean 0 and unknown variance σ_k^2 and that these variables are independent. The parameter σ_k^2 determines the degree of heterogeneity of the $p_{ij}^{(k)}$'s : great values of σ_k^2 imply high degree of heterogeneity.

Before we end this subsection, we will make some comments about the assumed models. First, the multinomial distribution of the observed M_i's (which is the one used in the likelihood function) implies that people are distributed independently and with equal probability on the sites of the sampling frame. This assumption is difficult to satisfy in actual situations; however, as will be shown later, the likelihood function depends on the observed M_i's basically through their sum M and since NM/n is a design-based estimator of τ_1, that is, it is a distribution free estimator, it follows that the MLE of τ_1 will be also robust to deviations from the multinomial distribution of the M_i's. Nevertheless, deviations from this model will affect the performance of variance estimators and confidence intervals derived under this assumption. Second, the Rasch model given by (3.2) implies the following: (i) the link-probability $p_{ij}^{(k)}$ depends only on two effects: the sociability of the people in cluster A_i and that of person $j \in U_k - A_i$; (ii) the two effects are additive, and (iii) for any site A_i in the frame and any person $j \in U - A_i$, $p_{ij}^{(k)} > 0$. Model (3.2) is a particular case of a generalized linear mixed model. (See Agresti 2002, Section 2.1, for a brief review of this type of model.) Therefore, we could incorporate the network structures of the people in cluster A_i and person $j \in U_k - A_i$ to model the link-probability $p_{ij}^{(k)}$ by extending model (3.2) to one that includes covariates associated with person j, with cluster A_i, and their interaction terms. However, if we used a more general model than (3.2), we would make the problem of inference much more difficult than that we face in this work. Thus, in spite of the relative simplicity of model (3.2), we expect that it still captures the heterogeneity of the link-probabilities and allow us to make inferences about the τ's at least at the correct order of magnitude.

3.2 Likelihood function

The easiest way of constructing the likelihood function is to factorize it into different components. One of them is associated with the probability of selecting the initial sample S_0, which is given by the multinomial distribution (3.1), that is,

$$L_{\text{MULT}}(\tau_1) \propto \frac{\tau_1!}{(\tau_1 - m)!}(1 - n/N)^{\tau_1 - m}.$$

Two other components are associated with the conditional probabilities of the configurations of links between the people in $U_k - S_0, k = 1, 2,$ and the clusters $A_i \in S_A$, given S_A. To derive these factors we need to compute the probabilities of some events. Let $\mathbf{X}_j^{(k)} = \left(X_{1j}^{(k)}, \ldots, X_{nj}^{(k)}\right)$ be the n – dimensional vector of link-indicator variables $X_{ij}^{(k)}$ associated with the j^{th} person in $U_k - S_0$. Notice that $\mathbf{X}_j^{(k)}$ indicates which clusters $A_i \in S_A$ are linked to that person. Let $\mathbf{x} = (x_1, \ldots, x_n)$ be a vector whose i^{th} element is 0 or $1, i = 1, \ldots, n$. Because of the assumptions we made about the distributions of the variables $X_{ij}^{(k)}$'s, we have that the conditional probability, given $\beta_j^{(k)}$, that $\mathbf{X}_j^{(k)}$ equals $\mathbf{x}$, that is, the probability that the j^{th} person is linked to only those clusters $A_i \in S_A$ such that the i^{th} element x_i of $\mathbf{x}$ equals 1, is

$$\Pr\left(\mathbf{X}_j^{(k)} = \mathbf{x} \,\middle|\, \beta_j^{(k)}, S_A\right) = \prod_{i=1}^{n}\left[p_{ij}^{(k)}\right]^{x_i}\left[1 - p_{ij}^{(k)}\right]^{1-x_i} = \prod_{i=1}^{n}\frac{\exp\left[x_i\left(\alpha_i^{(k)} + \beta_j^{(k)}\right)\right]}{1 + \exp\left(\alpha_i^{(k)} + \beta_j^{(k)}\right)}.$$

Therefore, the probability that the vector of link-indicator variables associated with a randomly selected person in $U_k - S_0$ equals $\mathbf{x}$ is

$$\pi_{\mathbf{x}}^{(k)}(\boldsymbol{\alpha}_k, \sigma_k) = \int\prod_{i=1}^{n}\frac{\exp\left[x_i\left(\alpha_i^{(k)} + \sigma_k z\right)\right]}{1 + \exp\left(\alpha_i^{(k)} + \sigma_k z\right)}\phi(z)\,dz,$$

where $\boldsymbol{\alpha}_k = \left(\alpha_1^{(k)}, \ldots, \alpha_n^{(k)}\right)$ and $\phi(\cdot)$ denotes the probability density function of the standard normal distribution $[\mathrm{N}(0,1)]$.

As in Coull and Agresti (1999), instead of using $\pi_{\mathbf{x}}^{(k)}(\boldsymbol{\alpha}_k, \sigma_k)$ in the likelihood function we will use its Gaussian quadrature approximation $\tilde{\pi}_{\mathbf{x}}^{(k)}(\boldsymbol{\alpha}_k, \sigma_k)$ given by

$$\tilde{\pi}_{\mathbf{x}}^{(k)}(\boldsymbol{\alpha}_k, \sigma_k) = \sum_{t=1}^{q}\prod_{i=1}^{n}\frac{\exp\left[x_i\left(\alpha_i^{(k)} + \sigma_k z_t\right)\right]}{1 + \exp\left(\alpha_i^{(k)} + \sigma_k z_t\right)}\nu_t, \tag{3.3}$$

where q is a fixed constant and $\{z_t\}$ and $\{\nu_t\}$ are obtained from tables.

We are now in conditions of computing the two above mentioned factors of the likelihood function. Let $\Omega = \{(x_1, \ldots, x_n) : x_i = 0,1; i = 1, \ldots, n\}$, the set of all n – dimensional vectors such that each one of their elements is 0 or 1. For $\mathbf{x} = (x_1, \ldots, x_n) \in \Omega$, let $R_{\mathbf{x}}^{(k)}$ be the random variable that indicates the number of distinct people in $U_k - S_0$ whose vectors of link-indicator variables are equal to $\mathbf{x}$. Finally, let R_k be the random variable that indicates the number of distinct people in $U_k - S_0$ that are linked to at least one cluster $A_i \in S_A$. Notice that $R_k = \sum_{\mathbf{x}\in\Omega-\{\mathbf{0}\}} R_{\mathbf{x}}^{(k)}$, where $\mathbf{0}$ denotes the n – dimensional vector of zeros.

Because of the assumptions we made about the distributions of the variables $X_{ij}^{(k)}$'s, we have that the conditional joint probability distribution of the variables $\{R_{\mathbf{x}}^{(1)}\}_{\mathbf{x}\in\Omega-\{\mathbf{0}\}}$ and $\tau_1 - m - R_1$, given that $\{M_i = m_i\}_{i=1}^{n}$, is a multinomial distribution with parameter of size $\tau_1 - m$ and probabilities $\{\pi_{\mathbf{x}}^{(1)}(\boldsymbol{\alpha}_1, \sigma_1)\}_{\mathbf{x}\in\Omega-\{\mathbf{0}\}}$ and $\pi_{\mathbf{0}}^{(1)}(\boldsymbol{\alpha}_1, \sigma_1)$, and that of the variables $\{R_{\mathbf{x}}^{(2)}\}_{\mathbf{x}\in\Omega-\{\mathbf{0}\}}$ and $\tau_2 - R_2$ is a multinomial distribution with parameter of size τ_2 and probabilities $\{\pi_{\mathbf{x}}^{(2)}(\boldsymbol{\alpha}_2, \sigma_2)\}_{\mathbf{x}\in\Omega-\{\mathbf{0}\}}$ and $\pi_{\mathbf{0}}^{(2)}(\boldsymbol{\alpha}_2, \sigma_2)$.

Therefore, the factors associated with the probabilities of the configurations of links between the people in $U_k - S_0, k = 1, 2,$ and the clusters $A_i \in S_A$ are

$$L_1(\tau_1, \boldsymbol{\alpha}_1, \sigma_1) \propto \frac{(\tau_1 - m)!}{(\tau_1 - m - r_1)!} \prod_{\mathbf{x}\in\Omega-\{\mathbf{0}\}} \left[\tilde{\pi}_{\mathbf{x}}^{(1)}(\boldsymbol{\alpha}_1, \sigma_1)\right]^{r_{\mathbf{x}}^{(1)}} \left[\tilde{\pi}_{\mathbf{0}}^{(1)}(\boldsymbol{\alpha}_1, \sigma_1)\right]^{\tau_1 - m - r_1}$$

and

$$L_2(\tau_2, \boldsymbol{\alpha}_2, \sigma_2) \propto \frac{\tau_2!}{(\tau_2 - r_2)!} \prod_{\mathbf{x}\in\Omega-\{\mathbf{0}\}} \left[\tilde{\pi}_{\mathbf{x}}^{(2)}(\boldsymbol{\alpha}_2, \sigma_2)\right]^{r_{\mathbf{x}}^{(2)}} \left[\tilde{\pi}_{\mathbf{0}}^{(2)}(\boldsymbol{\alpha}_2, \sigma_2)\right]^{\tau_2 - r_2}.$$

The last component of the likelihood function is associated with the conditional probability, given S_A, of the configuration of links between the people in S_0 and the clusters $A_i \in S_A$. To derive this factor firstly observe that by the definition of the indicator variables $X_{ij}^{(k)}$'s, the i^{th} element of the vector of link-indicator variables associated with a person in $A_i \in S_A$ is equal to zero. Thus, let $\Omega_{-i} = \{(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) : x_j = 0,1,\ j \neq i,\ j = 1, \ldots, n\}$, that is, the set of all $(n-1)$ – dimensional vectors obtained from the vectors in Ω by omitting their i^{th} coordinate. For $\mathbf{x} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \in \Omega_{-i}$ let $R_{\mathbf{x}}^{(A_i)}$ be the random variable that indicates the number of distinct people in $A_i \in S_A$ such that their vectors of link-indicator variables, when the i^{th} coordinate is omitted, equal $\mathbf{x}$. Finally, let $R^{(A_i)}$ be the random variable that indicates the number of distinct people in $A_i \in S_A$ that are linked to at least one site $A_j \in S_A, j \neq i$. Notice that $R^{(A_i)} = \sum_{\mathbf{x}\in\Omega_{-i}-\{\mathbf{0}\}} R_{\mathbf{x}}^{(A_i)}$, where $\mathbf{0}$ denotes the $(n-1)$ – dimensional vector of zeros. Then, as in the previous cases, the conditional joint probability distribution of the variables $\{R_{\mathbf{x}}^{(A_i)}\}_{\mathbf{x}\in\Omega_{-i}-\{\mathbf{0}\}}$ and $m_i - R^{(A_i)}$, given that $\{M_i = m_i\}_{i=1}^{n}$, is a multinomial distribution with parameter of size m_i and probabilities $\{\pi_{\mathbf{x}}^{(A_i)}(\boldsymbol{\alpha}_1^{(-i)}, \sigma_1)\}_{\mathbf{x}\in\Omega_{-i}-\{\mathbf{0}\}}$ and $\pi_{\mathbf{0}}^{(A_i)}(\boldsymbol{\alpha}_1^{(-i)}, \sigma_1)$, where $\boldsymbol{\alpha}_1^{(-i)} = (\alpha_1^{(1)}, \ldots, \alpha_{i-1}^{(1)}, \alpha_{i+1}^{(1)}, \ldots, \alpha_n^{(1)})$ and

$$\pi_{\mathbf{x}}^{(A_i)}\left(\boldsymbol{\alpha}_1^{(-i)},\sigma_1\right) = \int\prod_{j\neq i}^{n}\frac{\exp\left[x_j\left(\alpha_j^{(1)}+\sigma_1 z\right)\right]}{1+\exp\left(\alpha_j^{(1)}+\sigma_1 z\right)}\phi(z)\,dz.$$

Therefore, the probability of the configuration of links between the people in S_0 and the clusters $A_j \in S_A$ is given by the product of the previous multinomial probabilities (one for each $A_i \in S_A$), and consequently the factor of the likelihood associated with that probability is

$$L_0\left(\boldsymbol{\alpha}_1,\sigma_1\right) \propto \prod_{i=1}^{n}\prod_{\mathbf{x}\in\Omega_{-i}-\{\mathbf{0}\}}\left[\tilde{\pi}_{\mathbf{x}}^{(A_i)}\left(\boldsymbol{\alpha}_1^{(-i)},\sigma_1\right)\right]^{r_{\mathbf{x}}^{(A_i)}}\left[\tilde{\pi}_{\mathbf{0}}^{(A_i)}\left(\boldsymbol{\alpha}_1^{(-i)},\sigma_1\right)\right]^{m_i - r^{(A_i)}},$$

where

$$\tilde{\pi}_{\mathbf{x}}^{(A_i)}\left(\boldsymbol{\alpha}_1^{(-i)},\sigma_1\right) = \sum_{t=1}^{q}\prod_{j\neq i}^{n}\frac{\exp\left[x_j\left(\alpha_j^{(1)}+\sigma_1 z_t\right)\right]}{1+\exp\left(\alpha_j^{(1)}+\sigma_1 z_t\right)}\nu_t, \tag{3.4}$$

is the Gaussian quadrature approximation to the probability $\pi_{\mathbf{x}}^{(A_i)}\left(\boldsymbol{\alpha}_1^{(-i)},\sigma_1\right)$.

From the previous results we have that the likelihood function is given by

$$L\left(\tau_1,\tau_2,\alpha_1,\alpha_2,\sigma_1,\sigma_2\right) = L_{(1)}\left(\tau_1,\boldsymbol{\alpha}_1,\sigma_1\right)L_{(2)}\left(\tau_2,\boldsymbol{\alpha}_2,\sigma_2\right),$$

where

$$L_{(1)}\left(\tau_1,\boldsymbol{\alpha}_1,\sigma_1\right) = L_{\text{MULT}}\left(\tau_1\right)L_1\left(\tau_1,\boldsymbol{\alpha}_1,\sigma_1\right)L_0\left(\boldsymbol{\alpha}_1,\sigma_1\right)$$

and

$$L_{(2)}\left(\tau_2,\boldsymbol{\alpha}_2,\sigma_2\right) = L_2\left(\tau_2,\boldsymbol{\alpha}_2,\sigma_2\right).$$

In the comments at the end of Subsection 3.1 was indicated that the likelihood function depends on the M_i's basically through their sum M. This can be seen by noting that only the factor L_0 depends directly through the M_i's. The factors L_{MULT} and L_1 depend on the M_i's through M, whereas the factor $L_{(2)}$ does not depend on the M_i's.

3.3 Unconditional maximum likelihood estimators

Numerical maximization of the likelihood function $L\left(\tau_1,\tau_2,\alpha_1,\alpha_2,\sigma_1,\sigma_2\right)$ with respect to the parameters yields the ordinary or unconditional maximum likelihood estimators (UMLEs) $\hat{\tau}_k^{(U)},\hat{\alpha}_k^{(U)}$ and $\hat{\sigma}_k^{(U)}$ of τ_k,α_k and $\sigma_k, k = 1, 2$. Consequently the UMLE of $\tau = \tau_1 + \tau_2$ is $\hat{\tau}^{(U)} = \hat{\tau}_1^{(U)} + \hat{\tau}_2^{(U)}$. Closed forms for the UMLEs do not exist; however, using the asymptotic approximation $\partial\ln\left(\tau_k!\right)/\partial\tau_k \approx \ln\left(\tau_k\right)$, we get the following approximations to $\hat{\tau}_1^{(U)}$ and $\hat{\tau}_2^{(U)}$:

$$\hat{\tau}_1^{(U)} = \frac{M + R_1}{1-(1-n/N)\,\tilde{\pi}_{\mathbf{0}}^{(1)}\left(\hat{\boldsymbol{\alpha}}_1^{(U)},\hat{\sigma}_1^{(U)}\right)} \quad\text{and}\quad \hat{\tau}_2^{(U)} = \frac{R_2}{1-\tilde{\pi}_{\mathbf{0}}^{(2)}\left(\hat{\boldsymbol{\alpha}}_2^{(U)},\hat{\sigma}_2^{(U)}\right)}. \tag{3.5}$$

Notice that these expressions are not closed forms since $\hat{\alpha}_k^{(U)}$ and $\hat{\sigma}_k^{(U)}$ depend on $\hat{\tau}_k^{(U)}$, $k = 1, 2$. Nevertheless, these expressions are useful to get formulae for the asymptotic variances of $\hat{\tau}_1^{(U)}$ and $\hat{\tau}_2^{(U)}$.

3.4 Conditional maximum likelihood estimators

Another way to get MLEs of τ_k, α_k and σ_k is by using Sanathanan's (1972) approach, which yields conditional maximum likelihood estimators (CMLEs). These estimators are numerically simpler to compute than UMLEs. In addition, if covariates were used in the model for the link-probability $p_{ij}^{(k)}$, this approach could still be used to get estimators of τ_k, α_k and σ_k, whereas the unconditional likelihood approach could not since the values of the covariates associated with the non sampled elements would be unknown.

The idea in Sanathanan's approach is to factorize the pmf of the multinomial distributions of the frequencies $R_{\mathbf{x}}^{(k)}$ of the different configurations of links as follows:

$$\begin{aligned}
L_1(\tau_1, \boldsymbol{\alpha}_1, \sigma_1) &\propto f\left(\left\{r_{\mathbf{x}}^{(1)}\right\}_{\mathbf{x}\in\Omega-\{\mathbf{0}\}}, \tau_1 - m - r_1 \middle| \{m_i\}, \tau_1, \boldsymbol{\alpha}_1, \sigma_1\right) \\
&= f\left(\left\{r_{\mathbf{x}}^{(1)}\right\}_{\mathbf{x}\in\Omega-\{\mathbf{0}\}} \middle| \{m_i\}, \tau_1, r_1, \boldsymbol{\alpha}_1, \sigma_1\right) f\left(r_1 \middle| \{m_i\}, \tau_1, \boldsymbol{\alpha}_1, \sigma_1\right) \\
&\propto \prod_{\mathbf{x}\in\Omega-\{\mathbf{0}\}} \left[\frac{\tilde{\pi}_{\mathbf{x}}^{(1)}(\boldsymbol{\alpha}_1, \sigma_1)}{1 - \tilde{\pi}_{\mathbf{0}}^{(1)}(\boldsymbol{\alpha}_1, \sigma_1)}\right]^{r_{\mathbf{x}}^{(1)}} \times \frac{(\tau_1 - m)!}{(\tau_1 - m - r_1)!}\left[1 - \tilde{\pi}_{\mathbf{0}}^{(1)}(\boldsymbol{\alpha}_1, \sigma_1)\right]^{r_1} \left[\tilde{\pi}_{\mathbf{0}}^{(1)}(\boldsymbol{\alpha}_1, \sigma_1)\right]^{\tau_1 - m - r_1} \\
&= L_{11}(\boldsymbol{\alpha}_1, \sigma_1) L_{12}(\tau_1, \boldsymbol{\alpha}_1, \sigma_1)
\end{aligned}$$

and

$$\begin{aligned}
L_2(\tau_2, \boldsymbol{\alpha}_2, \sigma_2) &\propto f\left(\left\{r_{\mathbf{x}}^{(2)}\right\}_{\mathbf{x}\in\Omega-\{\mathbf{0}\}}, \tau_2 - r_2 \middle| \{m_i\}, \tau_2, \boldsymbol{\alpha}_2, \sigma_2\right) \\
&= f\left(\left\{r_{\mathbf{x}}^{(2)}\right\}_{\mathbf{x}\in\Omega-\{\mathbf{0}\}} \middle| \{m_i\}, \tau_2, r_2, \boldsymbol{\alpha}_2, \sigma_2\right) f\left(r_2 \middle| \{m_i\}, \tau_2, \boldsymbol{\alpha}_2, \sigma_2\right) \\
&\propto \prod_{\mathbf{x}\in\Omega-\{\mathbf{0}\}} \left[\frac{\tilde{\pi}_{\mathbf{x}}^{(2)}(\boldsymbol{\alpha}_2, \sigma_2)}{1 - \tilde{\pi}_{\mathbf{0}}^{(2)}(\boldsymbol{\alpha}_2, \sigma_2)}\right]^{r_{\mathbf{x}}^{(2)}} \times \frac{\tau_2!}{(\tau_2 - r_2)!}\left[1 - \tilde{\pi}_{\mathbf{0}}^{(2)}(\boldsymbol{\alpha}_2, \sigma_2)\right]^{r_2} \left[\tilde{\pi}_{\mathbf{0}}^{(2)}(\boldsymbol{\alpha}_2, \sigma_2)\right]^{\tau_2 - r_2} \\
&= L_{21}(\boldsymbol{\alpha}_2, \sigma_2) L_{22}(\tau_2, \boldsymbol{\alpha}_2, \sigma_2).
\end{aligned}$$

Observe that in each case the first factor $L_{k1}(\boldsymbol{\alpha}_k, \sigma_k)$ is proportional to the conditional joint pmf of the $\left\{R_{\mathbf{x}}^{(k)}\right\}_{\mathbf{x}\in\Omega-\mathbf{0}}$, given that $\{M_i = m_i\}_1^n$ and $R_k = r_k$, which is the multinomial distribution with parameter of size r_k and probabilities $\left\{\tilde{\pi}_{\mathbf{x}}^{(k)} \big/ \left[1 - \tilde{\pi}_{\mathbf{0}}^{(k)}\right]\right\}_{\mathbf{x}\in\Omega-\mathbf{0}}$, and that this distribution does not depend on τ_k. Notice also that the second factors $L_{12}(\tau_1, \boldsymbol{\alpha}_1, \sigma_1)$ and $L_{22}(\tau_2, \boldsymbol{\alpha}_2, \sigma_2)$ are proportional to the conditional pmfs of R_1 and R_2, given that $\{M_i = m_i\}_1^n$, which are the distributions $\text{Bin}\left(\tau_1 - m, 1 - \tilde{\pi}_{\mathbf{0}}^{(1)}\right)$ and $\text{Bin}\left(\tau_2, 1 - \tilde{\pi}_{\mathbf{0}}^{(2)}\right)$, respectively, where $\text{Bin}(\tau, \theta)$ denotes the Binomial distribution with parameter of size τ and probability θ.

The CMLEs $\hat{\boldsymbol{\alpha}}_k^{(C)}$ and $\hat{\sigma}_k^{(C)}$ of $\boldsymbol{\alpha}_k$ and $\sigma_k, k = 1, 2$ are obtained by maximizing numerically

$$L_{11}(\boldsymbol{\alpha}_1, \sigma_1) L_0(\boldsymbol{\alpha}_1, \sigma_1) \quad \text{and} \quad L_{21}(\boldsymbol{\alpha}_2, \sigma_2) \tag{3.6}$$

with respect to $(\boldsymbol{\alpha}_1, \sigma_1)$ and $(\boldsymbol{\alpha}_2, \sigma_2)$, respectively. Note that the factors in (3.6) do not depend on $\tau_k, k = 1, 2.$

Finally, by plugging the estimates $\hat{\boldsymbol{\alpha}}_k^{(C)}$ and $\hat{\sigma}_k^{(C)}$ into the factors of the likelihood function that depend on $\tau_k, k = 1, 2,$ and maximizing these factors, that is, maximizing $L_{12}\left(\tau_1, \hat{\boldsymbol{\alpha}}_1^{(C)}, \hat{\sigma}_1^{(C)}\right) L_{\text{MULT}}(\tau_1)$ and $L_{22}\left(\tau_2, \hat{\boldsymbol{\alpha}}_2^{(C)}, \hat{\sigma}_2^{(C)}\right)$, with respect to τ_1 and τ_2, respectively, we get that the CMLEs $\hat{\tau}_1^{(C)}$ and $\hat{\tau}_2^{(C)}$ of τ_1 and τ_2 are given by (3.5) but replacing $\hat{\boldsymbol{\alpha}}_k^{(U)}$ and $\hat{\sigma}_k^{(U)}$ by $\hat{\boldsymbol{\alpha}}_k^{(C)}$ and $\hat{\sigma}_k^{(C)}, k = 1, 2.$ Observe that these expressions for $\hat{\tau}_1^{(C)}$ and $\hat{\tau}_2^{(C)}$ are closed forms. The CMLE of τ is $\hat{\tau}^{(C)} = \hat{\tau}_1^{(C)} + \hat{\tau}_2^{(C)}$.

4 Confidence intervals

We will consider two types of confidence intervals (CIs) for the population sizes: profile likelihood and bootstrap CIs.

4.1 Profile likelihood confidence intervals

Several authors such as Cormack (1992), Evans, Kim and O'Brien (1996), Coull and Agresti (1999) and Gimenes, Choquet, Lamor, Scofield, Fletcher, Lebreton and Pradel (2005) have indicated that, in the context of capture-recapture sampling, profile likelihood confidence intervals (PLCIs) perform better than traditional Wald CIs when the sample size is not large. Some factors that affect the performance of Wald CIs are biases in the estimators of the population size, biases in the estimators of the variances and asymmetries in the distributions of the estimators of the population size. Besides, a Wald CI for the population size might present the drawback that its lower bound might be less than the number of captured elements. Notice that, with the exception of the first listed factor, none of the others affect the performance of PLCIs. Furthermore, Evans et al. (1996), based on Ratkowsky (1988), indicate that the nonlinear nature of the capture-recapture estimators is approximated by likelihood-based CIs better than by Wald CIs.

Since the proposed estimators resemble those used in capture-recapture sampling and based on the previous comments, we should expect that also in our case PLCIs performance better than Wald CIs. It is worth noting that if we wanted to use Wald CIs, we would need to compute estimators of the variances of the proposed estimators. One alternative is to construct estimators of their asymptotic variances by using Sanathanan's (1972) results; however, for n large, say 20 or greater, obtaining these type of estimator is computationally very expensive because for each estimator is required the construction of a $(n + 1) \times (n + 1)$ symmetric matrix whose elements are sums of 2^n terms.

To get PLCIs for τ_1, τ_2 and τ we will follow Coull and Agresti's (1999) approach. Thus, for fixed values τ_1, τ_2 and τ of the population sizes, let r_{10}, r_{20} and r_{00} be non-negative real numbers such that $\tau_1 = m + r_1 + r_{10}, \tau_2 = r_2 + r_{20}$ and $\tau = m + r_1 + r_2 + r_{00}$, where m, r_1 and r_2 are the observed values of the random variables M, R_1 and R_2. Then $100(1 - \alpha)\%$ PLCIs for τ_1, τ_2 and τ are defined as the following sets: $\{\tau_1 = m + r_1 + r_{10} : -2\ln[\Lambda_1(r_{10})] \leq \chi^2_{1,1-\alpha}\}$, $\{\tau_2 = r_2 + r_{20} : -2\ln[\Lambda_2(r_{20})] \leq \chi^2_{1,1-\alpha}\}$ and $\{\tau = m + r_1 + r_2 + r_{00} : -2\ln[\Lambda(r_{00})] \leq \chi^2_{1,1-\alpha}\}$, respectively, where

$$\begin{aligned}\Lambda_1(r_{10}) &= \max_{\alpha_1,\sigma_1} L_{(1)}(m + r_1 + r_{10}, \alpha_1, \sigma_1) / L_{(1)}(\hat{\tau}_1, \hat{\alpha}_1, \hat{\sigma}_1),\\ \Lambda_2(r_{20}) &= \max_{\alpha_2,\sigma_2} L_{(2)}(r_2 + r_{20}, \alpha_2, \sigma_2) / L_{(2)}(\hat{\tau}_2, \hat{\alpha}_2, \hat{\sigma}_2)\end{aligned}$$

and

$$\Lambda(r_{00}) = \max_{r_{10},\alpha_1,\alpha_2,\sigma_1,\sigma_2} L(m + r_1 + r_{10}, r_2 + r_{00} - r_{10}, \alpha_1, \alpha_2, \sigma_1, \sigma_2) / L(\hat{\tau}_1, \hat{\tau}_2, \hat{\alpha}_1, \hat{\alpha}_2, \hat{\sigma}_1, \hat{\sigma}_2),$$

$\hat{\tau}_k, \hat{\alpha}_k$ and $\hat{\sigma}_k$ are either the UMLEs or CMLEs of τ_k, α_k and $\sigma_k, k = 1, 2,$ and $\chi^2_{1,1-\alpha}$ is the $100(1-\alpha)^{\text{th}}$ quantile of the chi-square distribution with 1 degree of freedom.

Although PLCIs for τ_2 are not affected by a possible extra-Poisson variation of the M_i's [or strictly speaking extra-multinomial dispersion of $(M_1, \ldots, M_n)$] because they are obtained from the likelihood function $L_{(2)}(\tau_2, \boldsymbol{\alpha}_2, \sigma_2)$ which does not depend on these variables, we do not expect that the PLCIs for τ_1 and τ be robust to extra-Poisson variation of the M_i's; therefore we will consider adjusted PLCIs for τ_1 and τ that take into account this extra variation. Following the suggestion of Gimenes et al. (2005), the adjusted PLCIs are constructed as the previous ones but replacing the value $\chi^2_{1,1-\alpha}$ by the value $(s^2_M / \bar{m}) F_{1,n-1,1-\alpha}$, where $\bar{m} = m/n$ and $s^2_M = \sum_1^n (m_i - \bar{m})^2 / (n-1)$ are the sample mean and variance of the m_i's, and $F_{1,n-1,1-\alpha}$ is the $100(1-\alpha)^{\text{th}}$ quantile of the F distribution with 1 and $n-1$ degrees of freedom. Observe that $s^2_M / \bar{m}$ is obtained by dividing by $n-1$ the value of the Pearson chi-square test statistic to test the hypothesis that the conditional distribution of the observed M_i's, given that $\sum_1^n M_i = m$, is multinomial with parameter of size m and vector of probabilities $(1/n, \ldots, 1/n)$. The adjusted PLCIs should be used if the null hypothesis of the conditional multinomial distribution of the M_i's were rejected at the $100\alpha\%$ level of significance, that is, if $s^2_M / \bar{m} > \chi^2_{n-1,1-\alpha} / (n-1)$, otherwise the unadjusted PLCIs should be used.

It is worth noting that the calculation of PLCIs is a computationally expensive task; therefore, efficient numerical algorithms need to be used, such as the one proposed by Venzon and Moolgavkar (1988).

4.2 Bootstrap confidence intervals

We will present a variant of bootstrap to construct CIs for the population sizes τ_1, τ_2 and τ based on either the UMLEs or the CMLEs. The proposed variant is obtained by combining the bootstrap version for finite populations proposed by Booth, Butler and Hall (1994) and the parametric bootstrap variant (see Davison and Hinkley 1997, Chapter 2). This version of bootstrap is an extension of the one used by Félix-Medina and Monjardin (2006) in the case of homogeneous link-probabilities.

Since our proposed version of bootstrap is a parametric variant, we need to have estimates of all the parameters associated with the assumed models. Until now, the only parameters that have not yet been estimated are the random effects $\beta_j^{(k)}$'s. We will now derive a predictor of $\beta_j^{(k)}$. Thus, given the subset $U_k - S_0$ or $A_i \in S_A$ that contains the element j, the conditional joint pdf of $\mathbf{X}_j^{(k)}$ and $\beta_j^{(k)}$ is

$$\begin{aligned} f\left(\mathbf{x}_j^{(k)}, \beta_j^{(k)} \middle| j \in U_k - S_0, S_A\right) &= \Pr\left(\mathbf{X}_j^{(k)} = \mathbf{x}_j^{(k)} \middle| \beta_j^{(k)}, j \in U_k - S_0, S_A\right) f\left(\beta_j^{(k)}\right) \\ &\propto \prod_{i=1}^{n} \left[p_{ij}^{(k)}\right]^{x_{ij}^{(k)}} \left[1 - p_{ij}^{(k)}\right]^{1-x_{ij}^{(k)}} \exp\left[-\left(\beta_j^{(k)}\right)^2 \Big/ 2\sigma_k^2\right] \\ &\quad \text{if } j \in U_k - S_0,\ k = 1, 2, \end{aligned}$$

or

$$\begin{aligned} f\left(\mathbf{x}_j^{(1)}, \beta_j^{(1)} \middle| j \in A_{i'} \in S_A, S_A\right) &\propto \prod_{i \neq i'}^{n} \left[p_{ij}^{(1)}\right]^{x_{ij}^{(1)}} \left[1 - p_{ij}^{(1)}\right]^{1-x_{ij}^{(1)}} \exp\left[-\left(\beta_j^{(1)}\right)^2 \Big/ 2\sigma_1^2\right] \\ &\quad \text{if } j \in A_{i'} \in S_A,\ i' = 1, \ldots, n. \end{aligned}$$

We will use as a prediction or estimate of $\beta_j^{(k)}$ the value $\hat{\beta}_j^{(k)}$ that maximizes the conditional joint pdf of $\mathbf{X}_j^{(k)}$ and $\beta_j^{(k)}$ with the parameters α_k and σ_k set at either their UMLEs or their CMLEs. This procedure yields that $\hat{\beta}_j^{(k)}$ is given as the solution to the following equation:

$$\sum_{i=1}^{n} x_{ij}^{(k)} - \sum_{i=1}^{n} \frac{\exp\left[\hat{\alpha}_i^{(k)} + \beta_j^{(k)}\right]}{1 + \exp\left[\hat{\alpha}_i^{(k)} + \beta_j^{(k)}\right]} - \frac{1}{\hat{\sigma}_k^2} \beta_j^{(k)} = 0 \quad \text{if } j \in U_k - S_0,\ k = 1, 2,$$

or

$$\sum_{i \neq i'}^{n} x_{ij}^{(1)} - \sum_{i \neq i'}^{n} \frac{\exp\left[\hat{\alpha}_i^{(1)} + \beta_j^{(1)}\right]}{1 + \exp\left[\hat{\alpha}_i^{(1)} + \beta_j^{(1)}\right]} - \frac{1}{\hat{\sigma}_1^2} \beta_j^{(1)} = 0 \quad \text{if } \ j \in A_{i'} \in S_A,\ i' = 1, \ldots, n,$$

where $\hat{\alpha}_i^{(k)}$ and $\hat{\sigma}_k$ denote either the UMLEs or the CMLEs of $\alpha_i^{(k)}$ and $\sigma_k, i = 1, \ldots, n; k = 1, 2.$ Note that this equation implies that the predictor $\hat{\beta}_j^{(k)}$ of $\beta_j^{(k)}$ depends on the number of clusters that are linked to the element j, but not on the particular clusters to which that element is linked. Thus, if two persons j and j' in $U_k - S_0$ are linked to the same number of clusters in S_A, the predictors $\hat{\beta}_j^{(k)}$ and $\hat{\beta}_{j'}^{(k)}$ are equal one another. The same happens for two persons in $A_i \in S_A$.

Hereinafter, we will denote by $[\hat{\tau}_k]$, the nearest integer to $\hat{\tau}_k$, where $\hat{\tau}_k$ denotes either the UMLE or the CMLE of $\tau_k, k = 1, 2.$ The steps of the proposed bootstrap procedure are the following. (i) Construct a population vector $\mathbf{m}_{\text{Boot}}$ of N values of m_i's by repeating N/n times, assuming that N/n is an integer, the observed sample of n cluster sizes $\mathbf{m}_s = \{m_1, \ldots, m_n\}$. If N/n is not an integer, that is, if $N = an + b$, where a and $b, b < n$, are positive integers, then repeat a times $\mathbf{m}_s$ and add to this set a SRSWOR of b values of m_i's selected from $\mathbf{m}_s$. (ii) For each $k = 1, 2,$ construct a population vector $\hat{\alpha}_{\text{Boot}}^{(k)}$ of dimension N whose elements are the estimates $\hat{\alpha}_i^{(k)}$'s of the $\alpha_i^{(k)}$'s associated with the clusters whose sizes m_i's are in $\mathbf{m}_{\text{Boot}}$. (iii) Construct a population vector $\hat{\beta}_{\text{Boot}}^{(0)}$ whose elements are the estimates $\hat{\beta}_j^{(1)}$'s of the $\beta_j^{(1)}$'s associated with the people who belong to the clusters whose sizes m_i's are in $\mathbf{m}_{\text{Boot}}$. Observe that the dimension of this vector is not necessarily $[\hat{\tau}_1]$, but it equals the sum of the m_i's in $\mathbf{m}_{\text{Boot}}$. (iv) Construct a population vector $\hat{\beta}_{\text{Boot}}^{(1)}$ of dimension $[\hat{\tau}_1]$ whose first m elements are the estimates $\hat{\beta}_j^{(1)}$'s of the $\beta_j^{(1)}$'s associated with the people in S_0; the remaining $[\hat{\tau}_1] - m$ elements are the r_1 estimates $\hat{\beta}_j^{(1)}$'s of the $\beta_j^{(1)}$'s associated with the people in S_1 and the $[\hat{\tau}_1] - m - r_1$ estimates $\hat{\beta}_j^{(1)}$'s of the $\beta_j^{(1)}$'s associated with the non sampled people in U_1. These $[\hat{\tau}_1] - m$ elements $\hat{\beta}_j^{(1)}$'s are randomly placed after the first m elements $\hat{\beta}_j^{(1)}$ of $\hat{\beta}_{\text{Boot}}^{(1)}$. (v) Construct a population vector $\hat{\beta}_{\text{Boot}}^{(2)}$ of dimension $[\hat{\tau}_2]$ whose first r_2 elements are the estimates $\hat{\beta}_j^{(2)}$'s of the $\beta_j^{(2)}$'s associated with the people in S_2 and the remaining $[\hat{\tau}_2] - r_2$ elements are the estimates $\hat{\beta}_j^{(2)}$'s of the $\beta_j^{(2)}$'s associated with the non sampled people in U_2. (vi) Select a SRSWOR of n values m_i from $\mathbf{m}_{\text{Boot}}$. Let $S_A^{\text{Boot}} = \{i_1, \ldots, i_n\}$ be the set of indices of the m_i's in the sample. In addition, let $A_i^{\text{Boot}} = \left(\sum_{t=1}^{i-1} m_t, \sum_{t=1}^{i} m_t\right) \cap \mathbb{Z}$ be the set of indices j associated with the elements in the cluster whose index is $i \in S_A^{\text{Boot}}$, where m_t is the t^{th} element of $\mathbf{m}_{\text{Boot}}$ and $\mathbb{Z}$ is the set of the integer numbers. Finally, let $S_0^{\text{Boot}} = \bigcup_{i \in S_A^{\text{Boot}}} A_i^{\text{Boot}}$. (vii) For each $i \in S_A^{\text{Boot}}$ and $j \in \{1, \ldots, [\hat{\tau}_2]\}$ generate a value $x_{ij}^{(2)}$ by sampling from the Bernoulli distribution with mean $\hat{p}_{ij}^{(2)}$ given by (3.2), but replacing $\alpha_i^{(2)}$ and $\beta_j^{(2)}$ by their estimates $\hat{\alpha}_i^{(2)}$ and $\hat{\beta}_j^{(2)}$. Similarly, for each $i \in S_A^{\text{Boot}}$ and $j \in \{1, \ldots, [\hat{\tau}_1]\} - A_i^{\text{Boot}}$ generate a value $x_{ij}^{(1)}$ by sampling from the Bernoulli distribution with mean $\hat{p}_{ij}^{(1)}$, where the value of $\hat{\beta}_j^{(1)}$ that is used to compute $\hat{p}_{ij}^{(1)}$ is obtained from $\hat{\beta}_{\text{Boot}}^{(0)}$ if $j \in S_0^{\text{Boot}}$, and from $\hat{\beta}_{\text{Boot}}^{(1)}$ otherwise. (viii) Compute the estimates of τ_1, τ_2 and τ using the same procedure as that used to compute the original estimates $\hat{\tau}_1, \hat{\tau}_2$ and $\hat{\tau}$. (ix) Repeat the steps (vi)−(viii) a large enough number B of times. Let $\hat{\tau}_{1,b}^{\text{Boot}}, \hat{\tau}_{2,b}^{\text{Boot}}$ and $\hat{\tau}_b^{\text{Boot}}$ be the estimates obtained in the b^{th} bootstrap sample, $b = 1, \ldots, B$.

The final step of our proposed bootstrap variant consists in constructing the CIs for the population sizes. There exist several alternatives to do this. One is to construct them without assuming any distributions for the estimators $\hat{\tau}_1, \hat{\tau}_2$ and $\hat{\tau}$. As examples of this alternative are the basic and the percentile method. (See Davison and Hinkley 1997, Chapter 5, for descriptions of these methods.) In the basic method a $100(1 - \alpha)\%$ CI for τ is $\left[2\hat{\tau} - \hat{\tau}_{1-\alpha/2}^{\text{Boot}}, 2\hat{\tau} - \hat{\tau}_{\alpha/2}^{\text{Boot}}\right]$, and in the percentile method the CI is $\left[\hat{\tau}_{\alpha/2}^{\text{Boot}}, \hat{\tau}_{1-\alpha/2}^{\text{Boot}}\right]$, where $\hat{\tau}_{\alpha/2}^{\text{Boot}}$ and $\hat{\tau}_{1-\alpha/2}^{\text{Boot}}$ are the lower and upper $\alpha/2$ points of the empirical distribution obtained from $\hat{\tau}_b^{\text{Boot}}, b = 1, \ldots, B$. Although this type of alternative has good properties of robustness, it requires a large number B of bootstrap samples, say $B = 1{,}000$, and this might be a serious problem if $\hat{\tau}$ is costly to compute.

Another alternative to construct CIs is to assume a distribution for $\hat{\tau}$ and use the bootstrap sample to estimate the parameters of that distribution. In this case the number B of required bootstrap samples is not so large, say $50 \leq B \leq 200$ is generally enough. Examples of this alternative are the assumption that $\hat{\tau}$ is normally distributed and the one that $\hat{\tau} - \nu$ is lognormally distributed, where ν is the number of sampled elements. In the first case a $100(1 - \alpha)\%$ CI for τ is the well known Wald CI given by $\hat{\tau} \pm z_{\alpha/2}\sqrt{\hat{V}(\hat{\tau})}$, whereas in the second case the CI is $[\nu + (\hat{\tau} - \nu)/c,\ \nu + (\hat{\tau} - \nu) \times c]$, where $c = \exp\left\{z_{\alpha/2}\sqrt{\ln\left[1 + \hat{V}(\hat{\tau})/(\hat{\tau} - \nu)^2\right]}\right\}$, $z_{\alpha/2}$ is the upper $\alpha/2$ point of the standard normal distribution and $\hat{V}(\hat{\tau})$ is an estimate of the variance of $\hat{\tau}$. (See Williams, Nichols and Conroy 2002, Section 14.2, for a description of this type of CI.) It is worth noting that in the lognormal based CIs for τ_1, τ_2 and τ the values of ν are $m + r_1, r_2$ and $m + r_1 + r_2$, respectively.

An estimator $\hat{V}(\hat{\tau})$ of the variance of $\hat{\tau}$ could be computed using the sample variance of the bootstrap sample $\hat{\tau}_b^{\text{Boot}}, b = 1, \ldots, B.$ However, this estimator is not robust to extreme values of $\hat{\tau}_b^{\text{Boot}},$ which are likely to occur with the proposed estimators when the sampling rates are not large enough. Therefore, to use a robust estimator of $V(\hat{\tau})$ is a better strategy. One possibility is to use Huber's proposal 2 to jointly estimate the parameters of location and scale from the bootstrap sample. (See Staudte and Sheather 1990, Section 4.5, for a description of this method.) In particular, the estimate of the parameter of scale is an estimate of the standard deviation $\sqrt{\hat{V}(\hat{\tau})}$ of $\hat{\tau}.$

5 Sample size determination

We will present a procedure to determine the initial sample size $n.$ This procedure is based on stringent assumptions, but, as was indicated by one of the reviewers of the paper, it could nevertheless be very useful for researchers who want to apply this sampling design.

The first step is to compute the asymptotic variances of the proposed estimators. Although the variances depend on several unknown parameters, we can simplify them by assuming that the effects $\alpha_i^{(k)}$ of the sampled sites are homogeneous, that is, $\alpha_i^{(k)} = \alpha^{(k)}, i = 1, \ldots, n; k = 1, 2.$ Under this premise, the probabilities $\pi_{\mathbf{x}}^{(k)}$ and $\pi_{\mathbf{x}}^{(A_i)}$ that the vectors of link-indicator variables associated with randomly selected persons from $U_k - S_0$ and $S_0,$ respectively, equal $\mathbf{x}$ depend only on the number of 1's that appear in the vector $\mathbf{x}.$ Thus, their Gaussian quadrature approximations, given by (3.3) and (3.4), are simplified to

$$\tilde{\pi}_x^{(k)}\left(\boldsymbol{\theta}^{(k)}\right) = \tilde{\pi}_x^{(k)}\left(\alpha^{(k)}, \sigma_k\right) = \sum_{t=1}^{q} \frac{\exp\left[x\left(\alpha^{(k)} + \sigma_k z_t\right)\right]}{\left[1 + \exp\left(\alpha^{(k)} + \sigma_k z_t\right)\right]^n} \nu_t, \; x = 0, 1, \ldots, n,$$

and

$$\tilde{\pi}_x^{(A)}\left(\boldsymbol{\theta}^{(k)}\right) = \tilde{\pi}_x^{(A)}\left(\alpha^{(1)}, \sigma_1\right) = \sum_{t=1}^{q} \frac{\exp\left[x\left(\alpha^{(1)} + \sigma_1 z_t\right)\right]}{\left[1 + \exp\left(\alpha^{(1)} + \sigma_1 z_t\right)\right]^n} \nu_t, \; x = 0, 1, \ldots, n-1,$$

where $\boldsymbol{\theta}^{(k)} = \left(\theta_1^{(k)}, \theta_2^{(k)}\right) = \left(\alpha^{(k)}, \sigma_k\right).$

Following Sanathanan's (1972) procedure we get that the asymptotic variances of the proposed estimators are given by

$$V(\hat{\tau}_k) = \tau_k \Big/ \left(D_k - \mathbf{B}_k' \mathbf{A}_k^{-1} \mathbf{B}_k\right), \; k = 1, 2, \; \text{ and } \; V(\hat{\tau}) = V(\hat{\tau}_1) + V(\hat{\tau}_2),$$

where $\mathbf{A}_k = \left[a_{ij}^{(k)}\right]$ is a 2×2 matrix whose $a_{ij}^{(k)}$ element is

$$a_{ij}^{(k)} = \begin{cases} \left(1 - \dfrac{n}{N}\right) \displaystyle\sum_{x=0}^{n} \binom{n}{x} \dfrac{1}{\tilde{\pi}_x^{(1)}\left(\boldsymbol{\theta}^{(1)}\right)} \left[\dfrac{\partial \tilde{\pi}_x^{(1)}\left(\boldsymbol{\theta}^{(1)}\right)}{\partial \theta_i^{(1)}}\right] \left[\dfrac{\partial \tilde{\pi}_x^{(1)}\left(\boldsymbol{\theta}^{(1)}\right)}{\partial \theta_j^{(1)}}\right] & \\ + \dfrac{n}{N} \displaystyle\sum_{x=0}^{n-1} \binom{n-1}{x} \dfrac{1}{\tilde{\pi}_x^{(A)}\left(\boldsymbol{\theta}^{(1)}\right)} \left[\dfrac{\partial \tilde{\pi}_x^{(A)}\left(\boldsymbol{\theta}^{(1)}\right)}{\partial \theta_i^{(1)}}\right] \left[\dfrac{\partial \tilde{\pi}_x^{(A)}\left(\boldsymbol{\theta}^{(1)}\right)}{\partial \theta_j^{(1)}}\right] & \text{if } k = 1 \\ \displaystyle\sum_{x=0}^{n} \binom{n}{x} \dfrac{1}{\tilde{\pi}_x^{(2)}\left(\boldsymbol{\theta}^{(2)}\right)} \left[\dfrac{\partial \tilde{\pi}_x^{(2)}\left(\boldsymbol{\theta}^{(2)}\right)}{\partial \theta_i^{(2)}}\right] \left[\dfrac{\partial \tilde{\pi}_x^{(2)}\left(\boldsymbol{\theta}^{(2)}\right)}{\partial \theta_j^{(2)}}\right] & \text{if } k = 2, \end{cases}$$

$\mathbf{B}_k$ is a bi-dimensional vector whose elements are

$$b_i^{(k)} = -\left[\partial\tilde{\pi}_0^{(k)}\left(\boldsymbol{\theta}^{(k)}\right)\big/\partial\theta_i^{(k)}\right]\big/\tilde{\pi}_0^{(k)}(\boldsymbol{\theta}^{(k)}),\ i = 1, 2$$

and D_k is a real number given by

$$D_k = \begin{cases} \left[1-(1-n/N)\,\tilde{\pi}_0^{(1)}\left(\boldsymbol{\theta}^{(1)}\right)\right]\big/\left[(1-n/N)\,\tilde{\pi}_0^{(1)}\left(\boldsymbol{\theta}^{(1)}\right)\right] & \text{if } k = 1 \\ \left[1-\tilde{\pi}_0^{(2)}\left(\boldsymbol{\theta}^{(2)}\right)\right]\big/\tilde{\pi}_0^{(2)}\left(\boldsymbol{\theta}^{(2)}\right) & \text{if } k = 2. \end{cases}$$

It is worth noting that in the derivation of the asymptotic variances we have made the assumptions that $\tau_k \to \infty, k = 1, 2, m_i \to \infty, i = 1, \ldots, N,$ and N and n are fixed numbers.

To obtain numerical values of the variances we need to specify values for $\tau_k, \alpha^{(k)}, \sigma_k$ and $n.$ One way to do this is to assign values to τ_k, σ_k and to the proportion $\tilde{\pi}_{1+}^{(k)}$ of people in U_k who are linked at least to a particular site $A_i,$ which is common to all the sites and it is easier to specify than $\alpha^{(k)}.$ Then, for a given $n,$ the value of $\alpha^{(k)}$ is the solution to the equation

$$\tilde{\pi}_{1+}^{(k)} - \sum_{x=1}^{n}\binom{n-1}{x-1}\tilde{\pi}_x^{(k)}\left(\alpha^{(k)}, \sigma_k\right) = 0.$$

Once $\alpha^{(k)}, k = 1, 2,$ is obtained for a given $n,$ we can compute the numerical values of the variances $V(\hat{\tau}_k)$ and $V(\hat{\tau});$ the square roots of the relative variances $\sqrt{V(\hat{\tau}_k)/\tau_k^2}$ and $\sqrt{V(\hat{\tau})/\tau^2},$ and the sampling fractions $f_1 = 1-(1-n/N)\,\tilde{\pi}_0^{(1)},\ f_2 = 1-\tilde{\pi}_0^{(2)}$ and $f = (f_1 \times \tau_1 + f_2 \times \tau_2)/\tau.$ If the values of the square roots of the relative variances were not satisfactory, we could try different values of n until we get satisfactory values.

We have programmed this procedure in the R software programming language and it is available to the interested readers by requesting to the authors. To illustrate the procedure, let us suppose that we have a sampling frame of $N = 150$ sites and we assign the values $\tau_1 = 1{,}200,$ $\tau_2 = 400, \sigma_1 = \sigma_2 = 1,$ $\tilde{\pi}_{1+}^{(1)} = 0.05$ and $\tilde{\pi}_{1+}^{(2)} = 0.04,$ then for $n = 15$ we get that $V(\hat{\tau}_1) = 4{,}780.8,$ $V(\hat{\tau}_2) = 11{,}525.3,$ $V(\hat{\tau}) = 16{,}306.1,$ $\sqrt{V(\hat{\tau}_1)}/\tau_1 = 0.06,$ $\sqrt{V(\hat{\tau}_2)}/\tau_2 = 0.27,$ $\sqrt{V(\hat{\tau})}/\tau = 0.08,$ $f_1 = 0.50,$ $f_2 = 0.38$ and $f = 0.47.$

6 Monte Carlo studies

6.1 Populations constructed using artificial data

We constructed four artificial populations; a description of each one is presented in Table 6.1. Notice that in Populations I, III and IV the $N = 150$ values of the m_i's were obtained by sampling from a Poisson distribution, whereas in Population II by sampling from a zero truncated negative binomial distribution. In addition, in Populations I and II, the link-probabilities $p_{ij}^{(k)}$ were generated by the Rasch model (3.2). In Population III they were generated by that model but the random effects $\beta_j^{(k)}$ were obtained by sampling from a scaled Student's T distribution with six degrees of freedom and unit-variance instead of by sampling from the standard normal distribution. Finally, in Population IV, the $p_{ij}^{(k)}$'s were generated by the following latent class model proposed by Pledger (2000) in the context of capture-recapture studies: $p_{ij}^{(k)} = \exp\left[\mu^{(k)} + \alpha_i^{(k)} + \beta_j^{(k)} + (\alpha\beta)_{ij}^{(k)}\right] \Big/ \left\{1 + \exp\left[\mu^{(k)} + \alpha_i^{(k)} + \beta_j^{(k)} + (\alpha\beta)_{ij}^{(k)}\right]\right\}$, $i = 1, \ldots, n;\ j = 1, 2,$ and $k = 1, 2.$ In this model the people in U_k is divided into two latent classes $(j = 1, 2)$ according to their propensities to be linked to the sampled clusters. The probability that a randomly person in U_k is in class j is $p_j^{(k)}$ and the $p_{ij}^{(k)}$'s are the same for all the people in the class j.

The simulation experiment was carried out by repeatedly selecting r samples from each population by using the sampling design described in Section 2 with initial sample size $n = 15.$ Thus, each time that the value m_i was included in an initial sample, the value $x_{ij}^{(k)}$ was obtained by sampling from the Bernoulli distribution with mean $p_{ij}^{(k)}$. Because of the values assigned to n and to the parameters that appear in the expression of $p_{ij}^{(k)}$, the resulting sampling rates were $f_1 \approx 0.5$ and $f_2 \approx 0.4$. It is worth noting that the characteristics of the populations and samples considered in this study were not motivated by the ones of an actual study since this sampling design has not been applied yet. Thus, the populations and samples were constructed only with the purpose of analyzing the performance of the proposed point and interval estimators.

Table 6.1
Parameters of the simulated populations

Population I	Population II	Population III	Population IV
$N = 150$	$N = 150$	$N = 150$	$N = 150$
$M_i \sim$ Poisson	$M_i \sim$ Zero trunc. neg. binomial	$M_i \sim$ Poisson	$M_i \sim$ Poisson
$E(M_i) = 8$	$E(M_i) = 8$	$E(M_i) = 8$	$E(M_i) = 8$
$V(M_i) = 8$	$V(M_i) = 24$	$V(M_i) = 8$	$V(M_i) = 8$
$\tau_1 = 1{,}209$	$\tau_1 = 1{,}208$	$\tau_1 = 1{,}209$	$\tau_1 = 1{,}209$
$\tau_2 = 400$	$\tau_2 = 400$	$\tau_2 = 400$	$\tau_2 = 400$
$\tau = 1{,}609$	$\tau = 1{,}608$	$\tau = 1{,}609$	$\tau = 1{,}609$
$\alpha_i^{(k)} = \dfrac{c_k}{M_i^{1/4} + 0.001}$	$\alpha_i^{(k)} = \dfrac{c_k}{M_i^{1/4} + 0.001}$	$\alpha_i^{(k)} = \dfrac{c_k}{M_i^{1/4} + 0.001}$	$\alpha_i^{(k)} = \dfrac{-12}{M_i^{1/2} + 0.001}$
$c_1 = -5.45$	$c_1 = -5.45$	$c_1 = -5.45$	$\mu^{(1)} = -1.1;\ \mu^{(2)} = -1.2$
$c_2 = -5.85$	$c_2 = -5.85$	$c_2 = -5.85$	$\beta_1^{(k)} = 1.5;\ \beta_2^{(k)} = 0$
$\beta_j^{(k)} \sim N(0,1)$	$\beta_j^{(k)} \sim N(0,1)$	$\beta_j^{(k)} \sim \sqrt{2/3}\,T_6$	$(\alpha\beta)_{ij}^{(k)} \sim N\left(0, 1.25^2\right)$
			$p_1^{(k)} = 0.3 = 1 - p_2^{(k)}$

From each sample the following estimators of τ_1, τ_2 and τ were considered: the proposed UMLEs $\hat{\tau}_1^{(U)}, \hat{\tau}_2^{(U)}$ and $\hat{\tau}^{(U)}$; the proposed CMLEs $\hat{\tau}_1^{(C)}, \hat{\tau}_2^{(C)}$ and $\hat{\tau}^{(C)}$; the MLEs $\tilde{\tau}_1, \tilde{\tau}_2$ and $\tilde{\tau}$ proposed by Félix-Medina and Thompson (2004) and derived under the assumption of homogeneous link-probabilities, and the Bayesian-assisted estimators $\breve{\tau}_1, \breve{\tau}_2$ and $\breve{\tau}$ proposed by Félix-Medina and Monjardin (2006), derived also under the homogeneity assumption and using the following initial distributions for τ_1, τ_2 and $\alpha_i^{(k)} = \ln\left[p_i^{(k)} \big/ \left(1 - p_i^{(k)}\right)\right]$, where $p_i^{(k)}$ is given by (3.2), but setting $\beta_j^{(k)} = 0 : \xi(\tau_1 | \lambda_1) \propto (N\lambda_1)^{\tau_1} / \tau_1!$

and $\xi(\lambda_1) \propto \lambda_1^{a_1-1}\exp(-b_1\lambda_1)$; $\xi(\tau_2 | \lambda_2) \propto \lambda_2^{\tau_2} / \tau_2!$ and $\xi(\lambda_2) \propto \lambda_2^{a_2-1}\exp(-b_2\lambda_2)$, and $\xi\left(\alpha_i^{(k)} \middle| \theta_k\right) \propto \exp\left[-\left(\alpha_i^{(k)} - \theta_k\right)^2 / 2\sigma_k^2\right]$ and $\xi(\theta_k) \propto \exp\left[-(\theta_k - \mu_k)^2 / 2\gamma_k^2\right]$, where $a_1 = 1.0$, $b_1 = 0.1, a_2 = 6.0, b_2 = 0.01, \mu_k = -3.5$ and $\sigma_k^2 = \gamma_k^2 = 9.0$. These values assigned to the parameters of the initial distributions made them practically non-informative. The Gaussian quadrature approximations (3.3) and (3.4) to the probabilities $\pi_{\mathbf{x}}^{(k)}(\boldsymbol{\alpha}_k, \sigma_k)$ and $\pi_{\mathbf{x}}^{(A_i)}\left(\boldsymbol{\alpha}_1^{-i}, \sigma_1\right)$ were computed using $q = 40$ terms.

The performance of an estimator $\hat{\tau}$ of τ, say, was evaluated by means of its relative bias $(\mathrm{r-bias})$, the square root of its relative mean square error $\left(\sqrt{\mathrm{r-mse}}\right)$, and the medians of its relative estimation errors (mdre) and its absolute relative estimation errors (mdare) defined by $\mathrm{r-bias} = \sum_1^r (\hat{\tau}_i - \tau) / (r\tau)$, $\sqrt{\mathrm{r - mse}} = \sqrt{\sum_1^r (\hat{\tau}_i - \tau)^2 / \left(r\tau^2\right)}$, $\mathrm{mdre} = \mathrm{median}\{(\hat{\tau}_i - \tau)/\tau\}$ and $\mathrm{mdrae} = \mathrm{median}\{|(\hat{\tau}_i - \tau)/\tau|\}$, where $\hat{\tau}_i$ was the value of $\hat{\tau}$ obtained in the i^{th} sample, which in the case of the point estimators was 10,000.

We also considered the following 95% CIs for the τ's : the proposed PLCIs and adjusted for extra-Poisson variation PLCIs; the proposed bootstrap CIs based on $B = 100$ bootstrap samples and constructed assuming a lognormal distribution for $\hat{\tau} - \nu$ and estimating $\sqrt{\hat{V}(\hat{\tau})}$ by Huber's proposal 2 estimator of scale with tuning value $d = 1.5$; the design-based Wald CIs obtained from the MLEs $\tilde{\tau}_1, \tilde{\tau}_2$ and $\tilde{\tau}$ and proposed by Félix-Medina and Thompson (2004), and the design-based Wald CIs obtained from the Bayesian estimators $\breve{\tau}_1, \breve{\tau}_2$ and $\breve{\tau}$ and proposed by Félix-Medina and Monjardin (2006). It is worth noting that the PLCIs and adjusted PLCI were computed using the Venzon and Moolgavkar's (1988) method or an algorithm based on the definition of a PLCI when the first method failed to find the endpoints of the intervals.

The performance of a CI was evaluated by its coverage probability (cp), the mean of its relative lengths (mrl) and the median of its relative lengths (mdrl) defined as the proportion of samples in which the parameter was contained in the interval and the the mean and the median of the lengths of the intervals divided by the value of the parameter, respectively. Since carrying out a simulation study on the CIs using a large number of replicated samples is a very time consuming task, we evaluated the performance of the PLCIs for τ_1 and τ_2 using $r = 1{,}000$ samples; that of the PLCIs for τ using $r = 500$ samples and that of the bootstrap CIs using $r = 250$ samples. The performance of the CIs based on the estimators derived under the homogeneity assumption was evaluated by using $r = 10{,}000$ samples. The numerical study was carried out using the R software programming language [R Development Core Team (2013)].

The results of the study on the estimators of the population sizes are shown in Table 6.2 and in Figures 6.1 and 6.2. The main outcomes are the following. The distributions of the estimators UMLE $\hat{\tau}_1^{(U)}$ and CMLE $\hat{\tau}_1^{(C)}$ were more or less symmetrical about τ_1; thus, the two measures of bias $(\mathrm{r-bias}$ and mdre) showed similar values, as well as the two measures of variability $(\sqrt{\mathrm{r-mse}}$ and mdare). Both of these estimators performed acceptably well, except in Population III where the estimator $\hat{\tau}_1^{(U)}$ presented moderate problems of bias and $\hat{\tau}_1^{(C)}$ showed something more serious problems of bias. The distributions of the estimators $\hat{\tau}_2^{(U)}$ and $\hat{\tau}_2^{(C)}$ were skewed to the right with very long tails. This caused that the values of their $\mathrm{r-bias}$ and $\mathrm{r-mse}$ tended to be large. However, in terms of the medians of their relative errors (mdre), these estimators presented moderate problems of bias in Populations I and III and serious problems in Population IV. In terms of the medians of their absolute relative errors (mdare) these estimators showed moderate problems of instability in the first three populations and serious problems in the fourth population. The distributions of the estimators $\hat{\tau}^{(U)}$ and $\hat{\tau}^{(C)}$ were similar to those of the estimators $\hat{\tau}_2^{(U)}$ and $\hat{\tau}_2^{(C)}$; thus, the quantities $\mathrm{r-bias}$ and $\sqrt{\mathrm{r-mse}}$ were more sensitive to large values than the quantities mdre and mdare. Both of these estimators performed acceptably well in Populations I, II and IV; although in this last population the values of their $\sqrt{\mathrm{r-mse}}$ were large because of the reasons previously indicated. In Population III both estimators presented problems of bias.

Although the deviation from the assumed Poisson distribution of the M_i's increased the variability of all the proposed estimators, the increments were not large so that we consider that they have some robust properties against deviations from this assumption. The proposed estimators of τ_1 and τ were in addition robust to the deviation from the assumed Rasch model for the $p_{ij}^{(1)}$'s (although the values of the $\sqrt{\mathrm{r-mse}}$ of the estimators of τ were large, those of the median of their absolute relative errors were not). The deviation from the assumed normal distribution of the effects $\beta_j^{(k)}$ caused that all the proposed estimators presented problems of overestimation. Neither of the two types of proposed estimators UMLEs and CMLEs performed uniformly better than the other, but the UMLEs performed in a greater number of cases slightly better than the CMLEs.

Table 6.2
Relative biases, square roots of relative mean square errors and medians of relative errors and absolute relative errors of the estimators of the population sizes

Population		I				II				III				IV			
Sampling rates		f_1 0.51		f_2 0.40		f_1 0.50		f_2 0.40		f_1 0.51		f_2 0.40		f_1 0.51		f_2 0.40	
Estimator		r bias	$\sqrt{\mathrm{r\,mse}}$	mdre	mdare	r bias	$\sqrt{\mathrm{r\,mse}}$	mdre	mdare	r bias	$\sqrt{\mathrm{r\,mse}}$	mdre	mdare	r bias	$\sqrt{\mathrm{r\,mse}}$	mdre	mdare
Uncond.	$\hat{\tau}_1^{(U)}$	-0.01	0.06	-0.01	0.04	-0.00	0.08	-0.00	0.05	0.10	0.11	0.10	0.10	$-0.04_{6}^{0.20}$	0.08	-0.03	0.05
heter.	$\hat{\tau}_2^{(U)}$	-0.06	0.26	-0.11	0.17	$0.06^{0.02}$	0.35	-0.01	0.16	0.16	0.43	0.07	0.18	0.04_{2}^{15}	2.2	-0.19	0.25
MLEs	$\hat{\tau}^{(U)}$	-0.02	0.08	-0.03	0.05	$0.01^{0.02}$	0.10	0.01	0.06	0.11	0.15	0.10	0.10	-0.02_{7}^{15}	0.55	-0.6	0.08
Cond.	$\hat{\tau}_1^{(C)}$	-0.00	0.07	-0.01	0.05	0.01	0.07	0.00	0.05	0.18	0.19	0.17	0.17	$-0.05^{1.6}$	0.09	-0.05	0.07
heter.	$\hat{\tau}_2^{(C)}$	-0.04	0.26	-0.09	0.17	0.09	0.38	0.01	0.16	0.18	0.46	0.10	0.18	0.12_{2}^{21}	2.4	-0.14	0.23
MLEs	$\hat{\tau}^{(C)}$	-0.1	0.08	-0.02	0.05	0.03	0.11	0.01	0.06	0.18	0.22	0.16	0.16	-0.00_{2}^{23}	0.61	-0.06	0.08
Homo-	$\tilde{\tau}_1$	-0.28	0.28	-0.28	0.28	-0.31	0.31	-0.31	0.31	-0.30	0.30	-0.30	0.30	-0.18	0.19	-0.18	0.18
geneous	$\tilde{\tau}_2$	-0.40	0.40	-0.40	0.40	-0.40	0.40	-0.40	0.40	-0.40	0.40	-0.40	0.40	-0.30	0.32	-0.32	0.32
MLEs	$\tilde{\tau}$	-0.31	0.31	-0.31	0.31	-0.33	0.33	-0.33	0.33	-0.32	0.33	-0.32	0.32	-0.21	0.22	-0.21	0.21
Homo-	$\breve{\tau}_1$	-0.28	0.28	-0.28	0.28	-0.31	0.31	-0.31	0.31	-0.30	0.30	-0.30	0.30	-0.18	0.19	-0.18	0.18
geneous	$\breve{\tau}_2$	-0.39	0.39	-0.39	0.39	-0.39	0.40	-0.39	0.39	-0.39	0.40	-0.39	0.39	-0.27	0.30	-0.29	0.29
BEs	$\breve{\tau}$	-0.31	0.31	-0.31	0.31	-0.33	0.33	-0.33	0.33	-0.32	0.32	-0.32	0.32	-0.20	0.21	-0.20	0.20

Notes Results based on 10^4 samples. A superscript number indicates the percentage of samples in which the estimator was not computed because of numerical convergence problems and a subscript figure indicates the number of values of the estimator that exceeded 10^5 and that were not used to compute its $\mathrm{r-bias}$ and $\mathrm{r-mse}$. Upper bounds for the Monte Carlo errors of the estimates of the $\mathrm{r-bias}$ and the $\sqrt{\mathrm{r-mse}}$ of the estimators of the τ's were the following: τ_1 : 0.001 and 0.001; τ_2 : 0.004 and 0.011 in Pops. $\mathrm{I-III}$, and 0.027 and 0.39 in Pop. IV; τ : 0.001 and 0.002 in Pops. $\mathrm{I-III}$, and 0.007 and 0.095 in Pop. IV.

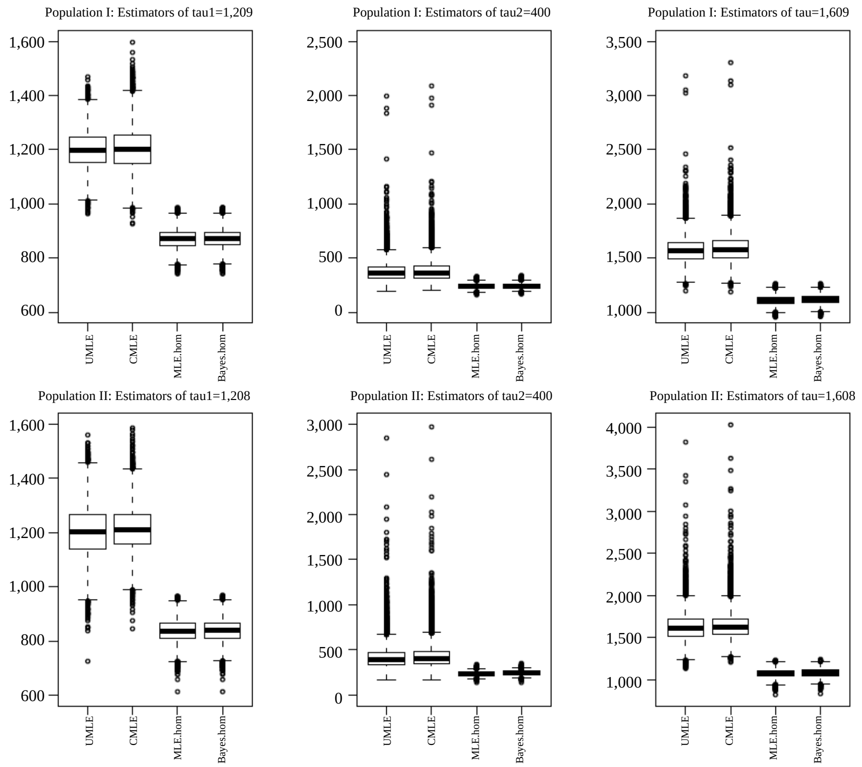
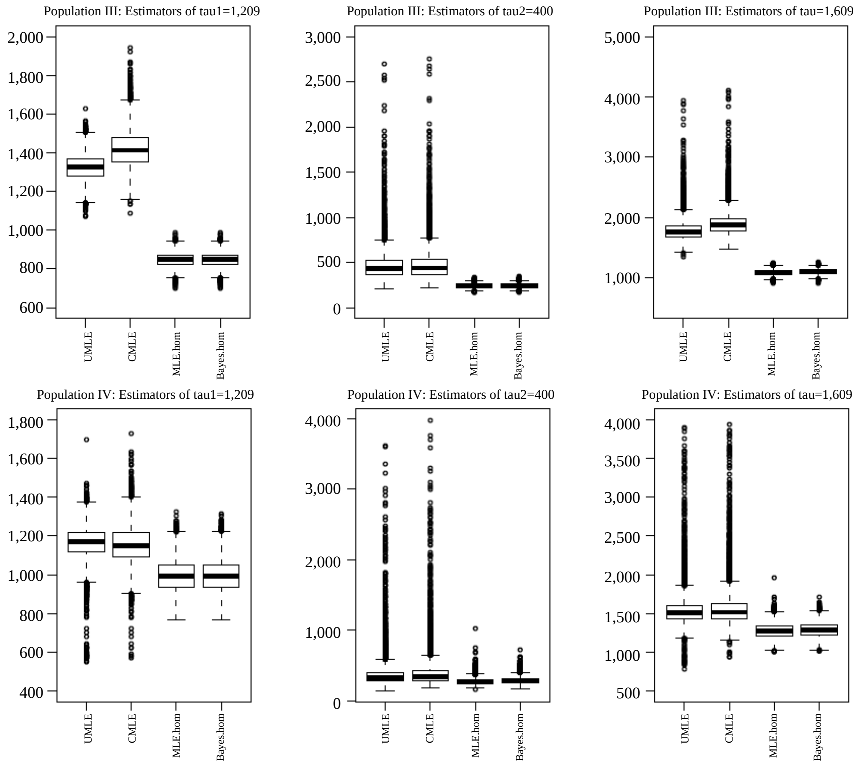


Figure 6.1 Boxplots for the values of the estimators of τ_1, τ_2 and τ in Populations I and II.



Note: 6 values of UMLE > 100,000 were omitted Note: 24 value of UMLE & CMLE > 4,000 were omitted Note: 36 & 34 values of UMLE & CMLE > 4,000 were omitted

Figure 6.2 Boxplots for the values of the estimators of τ_1, τ_2 and τ in Populations III and IV.

With regard to the estimators derived under the assumption of homogeneous $p_{ij}^{(k)}$, both the MLEs $\tilde{\tau}_1, \tilde{\tau}_2$ and $\tilde{\tau}$ and the Bayesian assisted estimators $\breve{\tau}_1, \breve{\tau}_2$ and $\breve{\tau}$ showed very similar behavior which was characterized by serious problems of bias that deteriorated their performance.

Notice that the percentages of samples in which the proposed estimators were not computed because of numerical convergence problems, as well as the number of samples in which the values of the estimators exceeded 10^5 and that not were used in calculating the reported results of $\mathrm{r-bias}$ and $\sqrt{\mathrm{r-mse}}$ because they would have been seriously affected, were not large, except in Population IV. As was indicated by a reviewer, computing the $\mathrm{r-bias}$ and $\mathrm{r-mse}$ of an estimator using only its available values lower than 10^5 favors the proposed estimators. We agree with that observation and for this reason we also reported measures of the performance of the estimators based on the medians of the relative errors and absolute relative errors which are robust to large values of the estimators. Thus, if we supposed that any time that an estimator was not computed its value had been very large and we computed the values of the measures of the performance of the estimators that are based on the medians using the complete set of observations the results would not have been different from those reported in Table 6.2, and our conclusions based on these measures would not have changed.

The results of the simulation study on the 95% CIs are shown in Table 6.3. The main outcomes are the following. All the PLCIs and adjusted PLCIs for τ_1 : the ones based on the UMLE $\hat{\tau}_1^{(U)}$ and those based on the CMLE $\hat{\tau}_1^{(C)}$, showed good values of the cp in Population I. The adjusted PLCIs presented also good values of the cp in Population II, but not the unadjusted PLCIs whose values of the cp were relatively low. In Population III the values of the cp of all the PLCIs and adjusted PLCIs for τ_1 were low, whereas in Population IV the values were only slightly low. A good characteristic of these CIs was that they showed pretty acceptable values of their mrl and mdrl in each of the situations that were considered. The PLCI for τ_2 based on $\hat{\tau}_2^{(U)}$ and the one based on $\hat{\tau}_2^{(C)}$ presented acceptable values of the cp in all the populations, except in Population IV, where the values were something low. However, in all the cases the mrl and mdrl of these CIs were so large that they were not useful for making reasonable inferences. Both PLCI for τ : the one based on the UMLE $\hat{\tau}^{(U)}$ and that based on the CMLE $\hat{\tau}^{(C)}$, performed acceptably well in Populations I and II, although the means of their relative lengths were large in Population II because this measure is not robust to great values of the lengths of the intervals. In the other populations these CIs showed problems of low coverage and/or large relative lengths; thus their performance was not good. Both types of adjusted PLCIs performed well only in Population I, in the other populations they presented large values of their relative lengths. Neither of the two types of CIs: the ones based on the UMLEs and those based on CMLEs performed uniformly better than the other, but those based on the UMLEs performed in a greater number of cases slightly better than those based on the CMLEs.

With respect to the bootstrap CIs, we have that each of the two types of CIs for τ_1 : the one based on $\hat{\tau}_1^{(U)}$ and that based on $\hat{\tau}_1^{(C)}$ performed well in Populations I, II and IV, although in this last population the values of their cp were slightly low. In Population III the values of their cp were very low because of the biases of the point estimators of τ_1. The two types of bootstrap CIs for τ_2 performed badly in all the populations because the values of their relative lengths were large. Finally, the two types of CIs for τ performed in general well. Notice that the values of their mrl tended to be large because this measure is not robust to great values of the lengths, whereas the values of their mdrl were acceptable. Neither of the two types of bootstrap CIs performed uniformly better than the other, but the CIs based on the UMLEs performed in most cases better than those based on the CMLEs.

Table 6.3
Coverage probabilities and means and medians of relative lengths of the 95% confidence intervals for the population sizes

Population	I			II			III			IV		
Sampling rates	f_1		f_2	f_1		f_2	f_1		f_2	f_1		f_2
	0.51		0.40	0.50		0.40	0.51		0.40	0.51		0.40
Conf. interval	cp	mrl	mdrl	cp	mrl	mdrl	cp	mrl	mdrl	cp	mrl	mdrl
PLCI- $\hat{\tau}_1^{(U)}$	0.95	0.22	0.22	0.85	0.22	0.22	0.61	0.24	0.24	$0.89^{1.6}$	0.23	0.24
Adj-PLCI- $\hat{\tau}_1^{(U)}$	0.94	0.24	0.24	0.98	0.42	0.41	0.69	0.27	0.26	$0.90^{1.6}$	0.25	0.26
PLCI- $\hat{\tau}_2^{(U)}$	0.94	1.4	0.98	0.95	2.8_2	1.3	0.95	2.6	1.6	0.77^{19}	7.1_6	1.4
PLCI- $\hat{\tau}^{(U)}$	0.95	0.66	0.59	0.97	0.91	0.65	1.0	1.0	0.79	0.86^{21}	2.1_1	0.65
Adj-PLCI- $\hat{\tau}^{(U)}$	0.92	0.75	0.62	$1.0^{7.0}$	5.8_{33}	2.1	$1.0^{0.20}$	1.7_1	0.87	0.90^{22}	2.7_4	0.78
Bootstr-CI- $\hat{\tau}_1^{(U)}$	0.94	0.23	0.23	0.94	0.33	0.31	0.59	0.25	0.25	$0.89^{0.40}$	0.24	0.25
Bootstr-CI- $\hat{\tau}_2^{(U)}$	0.87	1.4	0.86	0.97	3.9	1.6	0.97	4.7	2.1	0.83^{13}	4.6_3	0.90
Bootstr-CI- $\hat{\tau}^{(U)}$	0.93	0.38	0.28	0.99	0.97	0.49	0.98	1.1	0.52	0.89^{13}	1.2_3	0.31
PLCI- $\hat{\tau}_1^{(C)}$	0.96	0.24	0.23	0.91	0.25	0.24	0.76	0.31	0.29	$0.90^{2.1}$	0.25	0.25
Adj-PLCI- $\hat{\tau}_1^{(C)}$	0.95	0.26	0.25	0.99	0.44	0.42	0.81	0.33	0.32	$0.90^{2.1}$	0.27	0.27
PLCI- $\hat{\tau}_2^{(C)}$	0.94	1.4	0.98	0.95	2.7_2	1.3	0.95	2.5	1.6	0.85^{25}	7.5_4	1.7
PLCI- $\hat{\tau}^{(C)}$	0.95	0.64	0.54	$0.94^{1.6}$	1.2	0.62	$0.86^{2.0}$	3.1_1	0.74	0.93^{30}	2.9_1	0.84
Adj-PLCI- $\hat{\tau}^{(C)}$	0.96	0.82	0.59	$1.0^{7.6}$	7.2_{33}	2.6	$0.90^{3.8}$	3.6_6	1.2	0.94^{32}	2.9_5	0.90
Bootstr-CI- $\hat{\tau}_1^{(C)}$	0.96	0.29	0.29	0.98	0.31	0.31	0.48	0.40	0.39	$0.89^{1.6}$	0.31	0.30
Bootstr-CI- $\hat{\tau}_2^{(C)}$	0.90	1.5	1.0	0.98	4.7	1.7	0.97	5.5	2.4	0.94^{24}	3.4_6	1.3
Bootstr-CI- $\hat{\tau}^{(C)}$	0.95	0.44	0.34	1.0	1.1	0.49	0.96	1.3	0.62	0.94^{25}	0.88_6	0.43
Wald-CI- $\tilde{\tau}_1$	0.00	0.09	0.09	0.00	0.08	0.08	0.00	0.08	0.08	0.08	0.13	0.13
Wald-CI- $\tilde{\tau}_2$	0.00	0.17	0.16	0.00	0.17	0.16	0.00	0.16	0.16	0.18	0.34	0.31
Wald-CI- $\tilde{\tau}$	0.00	0.08	0.08	0.00	0.07	0.07	0.00	0.07	0.07	0.03	0.13	0.13
Wald-CI- $\breve{\tau}_1$	0.00	0.09	0.09	0.00	0.09	0.09	0.00	0.08	0.08	0.11	0.15	0.15
Wald-CI- $\breve{\tau}_2$	0.00	0.17	0.17	0.00	0.17	0.17	0.00	0.17	0.17	0.25	0.36	0.33
Wald-CI- $\breve{\tau}$	0.00	0.08	0.08	0.00	0.08	0.08	0.00	0.08	0.07	0.04	0.15	0.14

Notes A superscript number indicates the percentage of samples in which the confidence interval (CI) was not computed because of numerical convergence problems and a subscript figure indicates the number of values of the relative length of the CI that exceeded 10^5 and that were not used to compute its mrl. An upper bound (UB) for the Monte Carlo errors (MCEs) of the cps was 0.03. UBs for the MCEs of the estimates of the mrl were the following: PLCIs and adj. PLCIs for τ_1 : 0.003; PLCIs for τ_2 : 0.78; PLCIs and Adj. PLCIs for τ : 0.07 in Pop. I and 0.56 in Pops. II – IV; Bootstrap CIs for τ_1 : 0.005; Bootstrap CIs for τ_2 : 0.1 in Pop. I and 1.5 in Pops II – IV; Bootstrap CIs for τ : 0.02 in Pop. I and 0.37 in Pops. II – IV.

With regard to the CIs based on the point estimators derived under the homogeneity assumption, all of them showed null values of the cp, except in Population IV where the values were different from zero, but still too low. The bad performance of these CIs in terms of the cp was a result of the large biases of the point estimators. Thus, despite of the very small values of the $\mathrm{r-ml}$ of these intervals the very low values of their cp did not allow making reasonable inferences.

Observe that in the first three populations the percentages of samples in which the proposed CIs were not computed because of numerical convergence problems as well as the number of samples in which the values of their relative lengths exceeded 10^5 and that not were used in calculating the reported results of the mrl because they would have been seriously affected, were not large (less than 4%), except in the case of the adjusted PLCIs based on $\hat{\tau}^{(U)}$ and on $\hat{\tau}^{(C)}$ which in Population II were not computed in about 7% of the samples and the means of their relative lengths were computed without using 33 values greater than 10^5. However, in Population IV some percentages were close to 20% and others close to 30%. The large values of these percentages were, in part, consequence of the relatively large values of the percentages of samples in which the corresponding point estimators were not computed. It is clear that computing the measures of performance of a CI using only the cases in which the interval was obtained or computing the rml using only the samples in which the relative lengths were lower than 10^5 favors the proposed CIs. However, notice that practically in all cases in which those percentages were large, say larger than 5%, both the mrl and the mdrl of the intervals were large enough that those intervals were not useful for making inferences. Therefore, if the performance of these CIs was not good under this favorable assessment, it will not be good under a fairer evaluation. The exceptions to this pattern were the PLCI based on $\hat{\tau}_1^{(U)}$, and the two types of bootstrap CIs for τ which showed acceptable performance and large percentage of samples in which were not computed. So, the results of these CIs should be taken with reserve.

6.2 Population constructed using data from the Colorado Springs study on HIV/AIDS transmission

In this simulation study we constructed a population using data from the Colorado Springs study on heterosexual transmission of HIV/AIDS. As was indicated in the introduction to this paper, this epidemiological research was focused on a population of people who lived in the Colorado Springs metropolitan area from $1982-1992$ and who were at high risk of acquiring and transmitting HIV. That population included drug users, sex workers and their personal contacts, defined as those persons with whom they had close social, sexual or drug-associated relations. In that study, 595 initial responders were selected in a non-random fashion through a sexually transmitted disease clinic, a drug clinic, self-referral and street outreach. The responders were asked for a complete enumeration of their personal contacts and a total of 7,379 contacts who were not in the set of the initial responders were named and included in the study. In our simulation study the set U_1 was defined as the set of the 595 initial responders and, as in Félix-Medina and Monjardin (2010), they were grouped into $N = 105$ clusters of sizes m_i’s generated by sampling from a zero-truncated negative binomial distribution with parameter of size 2.5 and probability $2/3$. The sample mean and variance of the 105 values m_i's were 5.67 and 15.03, respectively. It is worth noting that most of the people who were assigned to the same cluster came from the same original source of recruitment. A person was defined to be linked to a cluster if he or she was a personal contact of at least one element in that cluster. Since, approximately 95% of the 7,379 contacts of the initial responders were linked to only one cluster, and this could affect the performance of the proposed estimators, in our study we defined the set U_2 as the subset of the 7,379 contacts formed by the 415 persons who were linked to at least two clusters plus the 379 sex workers who were linked to only one cluster. Thus, $\tau_1 = 595, \tau_2 = 794$ and $\tau = 1{,}389.$ It is worth noting that this population is the same as the one called "reduced population" by Félix-Medina and Monjardin (2010).

We set the sizes of the initial samples selected from the population to $n = 25.$ This value of n yielded the sampling rates: $f_1 = 0.46$ and $f_2 = 0.37.$ The simulation experiment was carried out as the previous one, except that each time that the value m_i was contained in an initial sample, all the people linked to cluster i were included in the sample. We used the same number of replications r and the same number B of bootstrap samples as those used in the previous study. In addition, the values of the parameters of the initial distributions that were used to construct the Bayesian-assisted estimators $\breve{\tau}_k$ and the value of q used to compute the Gaussian quadrature formulas (3.3) and (3.4) were the same as those used in the previous study.

The results of the simulation study are shown in Table 6.4. We can see that among the proposed estimators of the population sizes, only the estimators of τ_1 did not present problems of bias nor problems of instability. The estimators of τ_2 and τ exhibited serious problems of bias, particularly the estimators of τ_2, which affected their performance. As a result of the performance of the point estimators, only the adjusted PLCIs and bootstrap CIs for τ_1 performed acceptably well, although the values of the cp of the bootstrap CIs were slightly low. The unadjusted PLCIs for τ_1 showed low values of the cp because of the deviation from the assumed Poisson distribution of the M_i's. The PLCIs and bootstrap CIs for τ_2 and τ presented very large values of the mrl and mdrl that these intervals were not useful. Observe that the percentages of samples in which the proposed point and interval estimators of τ_1 were not computed because of numerical convergence problems were small (less than 1.2%). Therefore, they were virtually not favored by the evaluation procedure. In the case of the proposed point and intervals estimators of τ_2 and τ those percentages were large. However, if their performance was not good under this favorable assessment, it will not be good under a fairer evaluation.

With regard to the point estimators derived under the homogeneity assumption, we have that the MLEs $\tilde{\tau}_1$ and $\tilde{\tau}_2$ showed problems of bias which affected their performance; however, the estimator $\tilde{\tau}$ did not show problems of bias and its performance was acceptable. The small bias exhibited by this estimator might be explained by the fact that the negative bias of $\tilde{\tau}_1$ was canceled out by the positive bias of $\tilde{\tau}_2$. The Bayesian-assisted estimators performed similarly to the previous ones, although in this case the estimator $\breve{\tau}_2$ of τ_2 showed only mild problems of bias. The Wald CIs based on the MLEs and on the Bayesian-assisted estimators showed low values of the cp. However, since the values of the $\mathrm{r-ml}$ of these intervals were acceptable, the intervals for τ_2 and τ might provide some information about these parameters.

Table 6.4
Simulation results obtained for estimators and confidence intervals in a population constructed using data from the Colorado Springs study

		Point estimators				Confidence intervals			
Estimator		**$\mathrm{r-bias}$**	**$\sqrt{\mathrm{r-mse}}$**	**mdre**	**mdare**	**Conf. interval**	**cp**	**mdre**	**mdare**
						PLCI- $\hat{\tau}_1^{(U)}$	0.75	0.24	0.24
	$\hat{\tau}_1^{(U)}$	$-0.00_1^{0.03}$	0.10	-0.01	0.07	Adj-PLCI- $\hat{\tau}_1^{(U)}$	0.95	0.41	0.41
Uncond.						PLCI- $\hat{\tau}_2^{(U)}$	$0.39^{8.3}$	10_{18}	3.7
heter.	$\hat{\tau}_2^{(U)}$	$1.7_{16}^{3.5}$	4.5	0.79	0.79	PLCI- $\hat{\tau}^{(U)}$	$0.83^{8.6}$	5.7_6	2.1
MLEs						Adj-PLCI- $\hat{\tau}^{(U)}$	0.99^{21}	11_{47}	7.5
	$\hat{\tau}^{(U)}$	$0.95_{17}^{3.5}$	2.6	0.46	0.46	Bootstr-CI- $\hat{\tau}_1^{(U)}$	0.91	0.37	0.37
						Bootstr-CI- $\hat{\tau}_2^{(U)}$	$0.86^{3.2}$	11_{29}	3.6
						Bootstr-CI- $\hat{\tau}^{(U)}$	$0.88^{3.2}$	6.3_{28}	2.0
						PLCI- $\hat{\tau}_1^{(C)}$	$0.81^{1.2}$	0.30	0.27
	$\hat{\tau}_1^{(C)}$	$0.01^{0.57}$	0.12	0.01	0.08	Adj-PLCI- $\hat{\tau}_1^{(C)}$	$0.97^{1.2}$	0.45	0.44
Cond.						PLCI- $\hat{\tau}_2^{(C)}$	0.39^{10}	9.6_{17}	3.6
heter.	$\hat{\tau}_2^{(C)}$	$1.7_{10}^{4.7}$	4.5	0.80	0.80	PLCI- $\hat{\tau}^{(C)}$	0.89^{20}	6.2_9	2.6
MLEs						Adj-PLCI- $\hat{\tau}^{(C)}$	1.0^{27}	14_{32}	9.3
	$\hat{\tau}^{(C)}$	$0.96_{10}^{5.2}$	2.6	0.46	0.46	Bootstr-CI- $\hat{\tau}_1^{(C)}$	$0.86^{1.2}$	0.35	0.35
						Bootstr-CI- $\hat{\tau}_2^{(C)}$	$0.90^{8.0}$	9.7_{35}	3.9
						Bootstr-CI- $\hat{\tau}^{(C)}$	$0.91^{9.2}$	5.9_{34}	2.2
	$\tilde{\tau}_1$	-0.22	0.23	-0.22	0.22	Wald-CI- $\tilde{\tau}_1$	0.06	0.16	0.16
Homo-geneous	$\tilde{\tau}_2$	0.21	0.34	0.16	0.18	Wald-CI- $\tilde{\tau}_2$	0.71	0.60	0.53
MLEs	$\tilde{\tau}$	0.02	0.17	-0.00	0.10	Wald-CI- $\tilde{\tau}$	0.73	0.35	0.31
	$\breve{\tau}_1$	-0.22	0.23	-0.22	0.22	Wald-CI- $\breve{\tau}_1$	0.13	0.22	0.22
Homo-geneous	$\breve{\tau}_2$	0.12	0.22	0.10	0.13	Wald-CI- $\breve{\tau}_2$	0.72	0.45	0.43
BEs	$\breve{\tau}$	-0.02	0.13	-0.04	0.09	Wald-CI- $\breve{\tau}$	0.70	0.27	0.26

Notes A superscript number indicates the percentage of samples in which the estimator or confidence interval (CI) was not computed because of numerical convergence problems. A subscript figure indicates the number of values of the estimator or the relative length of a CI that exceeded 10^5 and that were not used to compute the $\mathrm{r-bias}$ and $\sqrt{\mathrm{r-mse}}$ of the estimator or the mrl of the CI. Upper bounds (UBs) for the Monte Carlo errors (MCEs) of the estimates of $\mathrm{r-bias}$ and $\sqrt{\mathrm{r-mse}}$ of the estimators of the τ's were the following: τ_1: 0.001 and 0.001; τ_2 : 0.05 and 0.31, and τ : 0.03 and 0.18. An UB for the MCEs of the estimates of cp was 0.02. UBs for the MCEs of the estimates of the mrl were the following: PLCIs and adj. PLCIs for τ_1 : 0.003; PLCIs for τ_2 and τ : 0.59; adj PLCIs for τ : 0.93; Bootstrap CIs for τ_1, τ_2 and τ : 0.005, 1.5 and 0.88, respectively.

7 Conclusions and suggestions for future research

The results of the simulation studies carried out in this research indicate that the two proposed estimators of τ_1 : $\hat{\tau}_1^{(U)}$ and $\hat{\tau}_1^{(C)}$, perform reasonably well in different situations. This evidences their robustness to several types of deviations from the assumed model. (Although $\hat{\tau}_1^{(C)}$ seems to be sensitive to deviations from the assumed normal distribution of the $\beta_j^{(k)}$'s.) On the other hand, the two proposed estimators of τ_2 : $\hat{\tau}_2^{(U)}$ and $\hat{\tau}_2^{(C)}$, present problems of bias and especially problems of instability if the sampling fraction in U_2 is not large enough, say it is not larger than 50%. In addition, small sampling fractions along with deviations from the assumed model for the link-probabilities increase the risk of numerical convergence problems. The two proposed estimators of τ : $\hat{\tau}^{(U)}$ and $\hat{\tau}^{(C)}$, perform similarly to the estimators $\hat{\tau}_1^{(U)}$ and $\hat{\tau}_1^{(C)}$ if τ_1 is much greater than τ_2 (as in the case of the artificial populations), perform similarly to the estimators $\hat{\tau}_2^{(U)}$ and $\hat{\tau}_2^{(C)}$ if τ_2 is much greater than τ_1 (as in the case of the Colorado Springs population), and perform as a combination of the performance of the estimators of τ_1 and τ_2 if the values of these parameters are not very different from each other. Finally, the estimators derived under the assumption of homogeneous link-probabilities present serious problems of bias if this assumption is not satisfied.

It is worth noting that our conclusion about the proposed estimators of τ_2 is based on the results of several small simulation studies that we carried out using sampling fractions greater than those used in the Monte Carlo studies reported in this paper. In one study carried out with the artificial populations, we increased the values of the link-probabilities $p_{ij}^{(k)}$ so that their average values were $\bar{p}_{ij}^{(1)} \approx 0.088$ and $\bar{p}_{ij}^{(2)} \approx 0.071$ and kept the sizes of the initial samples at $n = 15$. These changes yielded the sampling fractions $f_1 \approx 0.65$ and $f_2 \approx 0.55$. In another study also with the artificial populations, we reduced the values of the $p_{ij}^{(k)}$ so that $\bar{p}_{ij}^{(1)} = \bar{p}_{ij}^{(2)} \approx 0.016$ and increased the sizes of the initial samples to $n = 78$ in Populations $\mathrm{I}-\mathrm{III}$ and to $n = 67$ in Population IV. These changes yielded the sampling fractions $f_1 \approx 0.78$ and $f_2 \approx 0.55$. In both studies the estimators $\hat{\tau}_2^{(U)}$ and $\hat{\tau}_2^{(C)}$ performed acceptably well. (The results are not shown.) These outcomes indicate that these estimators also seem to have properties of robustness to deviations from the assumed models provided that large sampling fractions be used.

However, in a study with the Colorado Springs population using initial samples of sizes $n = 42$ which yielded sampling fractions $f_1 \approx 0.64$ and $f_2 \approx 0.56$, the estimators $\hat{\tau}_2^{(U)}$ and $\hat{\tau}_2^{(C)}$ presented serious problems of bias ($\mathrm{r-bias} \approx 1.0$ and $\mathrm{mdre} \approx 0.85$) which affected the values of their $\mathrm{r-mse}$ $\left(\sqrt{\mathrm{r-mse}} \approx 1.0\right)$ and $\mathrm{mdare}\left(\mathrm{mdare} \approx 0.85\right)$. Why these estimators did not perform well even with large sampling fractions? We think that the bad performance of these estimators is consequence of the very small average value of the $p_{ij}^{(2)}$'s $\left(\bar{p}_{ij}^{(2)} \approx 0.018\right)$ and the way the Monte Carlo studies were carried out. To clarify this statement, note the following. When $\bar{p}_{ij}^{(2)}$ is very small, say less than 0.02, the expected number of elements in U_2 that are linked to at least one site in the frame is much less than τ_2. For instance, in Population I when $\bar{p}_{ij}^{(2)} = 0.015$, this expected number was about $300 < 400 = \tau_2$. Therefore, if the sampling fraction f_2 is large enough, the estimates $\hat{\tau}_2^{(U)}$ and $\hat{\tau}_2^{(C)}$ will be close to τ_2 and will be much greater than the expected number of elements linked to at least one site in the frame. Thus, if we supposed that the Colorado Springs population was generated by a random process, then the 794 contacts linked to at least one site in the frame, and which we used as the value of τ_2, would be a much smaller value than the actual size of U_2. Consequently, the performance of $\hat{\tau}_2^{(U)}$ and $\hat{\tau}_2^{(C)}$ as estimators of the assumed value 794 of τ_2 would be very bad. This explanation was suggested and confirmed by the results of a small simulation study in which we considered Population I (in which every one of the assumptions is satisfied), but instead of carrying the study as is described in Subsection 5.1, we generated the complete set of values $x_{ij}^{(2)}$ of $X_{ij}^{(2)}$ by sampling from the Bernoulli distributions with means $p_{ij}^{(2)}, i = 1, \ldots, N; j = 1, \ldots, 400$ and we kept them fixed. Then, we defined the value of τ_2 as the number of elements of U_2 linked to at least one site in the frame. We considered two cases: large values of $p_{ij}^{(2)} \left(\bar{p}_{ij}^{(2)} = 0.071 \right)$ and small values $\left(\bar{p}_{ij}^{(2)} = 0.015 \right)$. In the first case $\tau_2 = 388,$ whereas in the second one $\tau_2 = 300.$ To have comparable results the sizes of the initial samples were set to $n = 15$ in the first case and to $n = 78$ in the second case, so that in both cases the number of sampled elements were about 220. The results of the numerical study showed that in the case of large values of $p_{ij}^{(2)}$ the estimators $\hat{\tau}_2^{(U)}$ and $\hat{\tau}_2^{(C)}$ performed well (because $388 \approx 400$), whereas in the case of small values of $p_{ij}^{(2)}$ these estimators performed badly (because $300 << 400$). We think that the results obtained in the last case are illustrative and explain the ones obtained in the Colorado Springs population.

With respect to the two types of proposed CIs: profile likelihood and bootstrap CIs, we can conclude that they need larger sample sizes than the point estimators to perform reasonably well. They are more sensitive to deviations from the assumed models than the point estimators. In addition, if small sampling fractions are used and deviations from the assumed model for the $p_{ij}^{(k)}$ are present, the occurrence of numerical convergence problems will be greater than in the case of point estimators.

From the previous observations we can conclude that in actual applications of this sampling methodology, a good strategy is to construct a sampling frame that covers the largest possible portion of the target population. This way, τ_1 would be close to $\tau,$ and the estimates of τ_1 could be used as estimates of $\tau.$ The advantage of this strategy is that the estimators $\hat{\tau}_1^{(U)}$ and $\hat{\tau}_1^{(C)}$ perform better than the estimators $\hat{\tau}_2^{(U)}$ and $\hat{\tau}_2^{(C)}$ because the first ones incorporate the information about the cluster sizes $m_i.$ Furthermore, this strategy makes possible to use the design-based estimator $N \sum_1^n m_i / n$ as an estimator of $\tau.$ The other factor that must be taken into account to have good estimates is to use large sampling fractions, say larger than 0.5. This suggestion is in agreement with the result reported by Xi, Watson and Yip (2008), who in the context of capture-recapture studies indicate that in presence of heterogeneous capture probabilities, a population size between 300 and 500, and a number of sampling occasions between 10 and 20, the minimum sampling fraction to have reliable estimates is at least 60%. Since the estimation of τ_2 is basically the same problem as that of estimating the population size in a capture-recapture study, we think that this conclusion also applies to our situation. In this line, we have developed a method to determine the size of the initial sample in order to have desired values of $\sqrt{V(\hat{\tau}_k) / \tau_k^2}, k = 1, 2$ and $\sqrt{V(\hat{\tau}) / \tau^2}.$ Although this procedure is based on stringent assumptions such as the homogeneity of the effects $\alpha_i^{(k)}$'s associated with the sites and the necessity of large values of the m_i's, the results seem to be satisfactory. For instance, the situation illustrated at the end of Section 5 correspond to that of the artificial populations considered in the Monte Carlo study and we can see that the results obtained by our procedure are very close to those reported for Populations I and II (Table 6.2), where the estimators of τ_2 and τ performed acceptably.

Finally, despite the drawbacks of the proposed point and interval estimators, they are a better alternative for making inferences about the population size than those based on the assumption of homogeneous link-probabilities. Obviously, our proposal need to be improved. The two major problems that need to be considered in future research are the instability of the estimators of τ_2 when the sampling fraction is not large enough and the not satisfactory performance of the confidence intervals. A possible solution to both problems is to use the Bayesian approach to construct estimators that incorporate information prior to sampling that the researcher has about the parameters. The point and interval estimators obtained by this approach might be more stable than those proposed in this paper because of the additional information used to construct them. Other possible solution to the problem of lack of robustness of the confidence intervals is to replace the assumption of the Poisson distribution of the M_i's by a more flexible distribution such as the negative binomial, and the assumption of the normal distribution of the effects $\beta_j^{(k)}$ by one of the distributions ordinarily used to increase the robustness of the estimators such as a mixture of normal distributions or a Student's T distribution.

Acknowledgements

This research was supported by grant PROFAPI 2008/054 of the *Universidad Autónoma de Sinaloa* and grant APOY-COMPL-2008/89777 of the *Consejo Nacional de Ciencia y Tecnología*, Mexico. We thank John Potterat and Steve Muth for allowing us to use the data from the Colorado Springs study. We also thank the reviewers for their helpful suggestions and comments which improved this work.

References

Agresti, A. (2002). *Categorical Data Analysis, Second edition*. New York: John Wiley & Sons, Inc.

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

Cormack, R.M. (1992). Interval estimation for mark-recapture studies of closed populations. *Biometrics,* 48, 567-576.

Coull, B.A., and Agresti, A. (1999). The use of mixed logit models to reflect heterogeneity in capture-recapture studies. *Biometrics,* 55, 294-301.

Dávid, B., and Snijders, T.A.B. (2002). Estimating the size of the homeless population in Budapest, Hungary. *Quality & Quantity,* 36, 291-303.

Davison, A.C., and Hinkley, D.V. (1997). *Bootstrap Methods and their Applications*. New York: Cambridge University Press.

Evans, M.A., Kim, H.-M. and O’Brien, T.E. (1996). An application of profile-likelihood based confidence interval to capture-recapture estimators. *Journal of Agricultural, Biological and Environmental Statistics,* 1, 131-140.

Félix-Medina, M.H., and Monjardin, P.E. (2006). Combining link-tracing sampling and cluster sampling to estimate the size of hidden populations: A bayesian-assisted approach. *Survey Methodology,* 32, 2, 187-195.

Félix-Medina, M.H., and Monjardin, P.E. (2010). Combining link-tracing sampling and cluster sampling to estimate totals and means of hidden human populations. *Journal of Official Statistics,* 26, 603-631.

Félix-Medina, M.H., and Thompson, S.K. (2004). Combining cluster sampling and link-tracing sampling to estimate the size of hidden populations. *Journal of Official Statistics,* 20, 19-38.

Frank, O., and Snijders, T.A.B. (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics,* 10, 53-67.

Gimenes, O., Choquet, R., Lamor, R., Scofield, P., Fletcher, D., Lebreton, J.-D. and Pradel, R. (2005). Efficient profile-likelihood confidence intervals for capture-recapture models. *Journal of Agricultural, Biological and Environmental Statistics,* 10, 1-13.

Heckathorn, D.D. (2002). Respondent-driven sampling II: deriving valid population estimates from chain-referral samples of hidden populations. *Social Problems,* 49, 11-34.

Johnston, L.G., and Sabin, K. (2010). Sampling hard-to-reach populations with respondent driven sampling. *Methodological Innovations Online,* 5, 2, 38-48.

Kalton, G. (2009). Methods for oversampling rare populations in social surveys. *Survey Methodology,* 35, 2, 125-141.

Karon, J.M., and Wejnert, C. (2012). Statistical methods for the analysis of time-location sampling data. *Journal of Urban Health,* 89, 565-586.

MacKellar, D., Valleroy, L., Karon, J., Lemp, G. and Janssen, R. (1996). The young men’s survey: Methods for estimating HIV sero-prevalence and risk factors among young men who have sex with men. *Public Health Reports,* 111, supplement 1, 138-144.

Magnani, R., Sabin, K., Saidel, T. and Heckathorn, D. (2005). Review of sampling hard-to-reach populations for HIV surveillance. *AIDS,* 19, S67-S72.

McKenzie, D.J., and Mistiaen, J. (2009). Surveying migrant households: a comparison of census-based, snowball and intercept point surveys. *Journal of the Royal Statistical Society, Series A,* 172, 339-360.

Munhib, F.B., Lin, L.S., Stueve, A., Miller, R.L., Ford, W.L., Johnson, W.D. and Smith, P. (2001). A venue-based method for sampling hard-to-reach populations. *Public Health Reports,* 116, supplement 1, 216-222.

Pledger, S. (2000). Unified maximum likelihood estimates for closed capture-recapture models using mixtures. *Biometrics*, 56, 434-442.

Potterat, J.J., Woodhouse, D.E., Muth, S.Q., Rothenberg, R.B., Darrow, W.W., Klovdahl, A.S. and Muth, J.B. (2004). Network dynamism: History and lessons of the Colorado Springs study. In *Network Epidemiology: A Handbook for Survey Design and Data Collection,* (Ed., M. Morris), New York: Oxford University Press, 87-114.

Potterat, J.J., Woodhouse, D.E., Rothenberg, R.B., Muth, S.Q., Darrow, W.W., Muth, J.B. and Reynolds, J.U. (1993). AIDS in Colorado Springs: Is there an epidemic? *AIDS,* 7, 1517-1521.

R Development Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ratkowsky, D.A. (1988). *Handbook of Nonlinear Regression Models*. New York: Marcel Dekker.

Rothenberg, R.B., Woodhouse, D.E., Potterat, J.J., Muth, S.Q., Darrow, W.W. and Klovdahl, A.S. (1995). Social networks in disease transmission: The Colorado Springs study. In *Social Networks, Drug Abuse, and HIV Transmission,* (Eds., R.H. Needle, S.G. Genser and R.T. II Trotter) NIDA Research Monograph 151. Rockville, MD: National Institute of Drug Abuse, 3-19.

Sanathanan, L. (1972). Estimating the size of a multinomial population. *Annals of Mathematical Statistics*, 43, 142-152.

Semaan, S. (2010). Time-space sampling and respondent-driven sampling with hard-to-reach populations. *Methodological Innovations Online,* 5, 2, 60-75.

Spreen, M. (1992). Rare populations, hidden populations, and link-tracing designs: What and why? *Bulletin de Méthodologie Sociologique,* 36, 34-58.

Staudte, R.G., and Sheather, S.J. (1990). *Robust Estimation and Testing*. New York: John Wiley & Sons, Inc.

Thompson, S.K., and Frank, O. (2000). Model-based estimation with link-tracing sampling designs. *Survey Methodology*, 26, 1, 87-98.

Venzon, D.J., and Moolgavkar, S.H. (1988). A method for computing profile-likelihood-based confidence intervals. *Applied Statistics*, 37, 87-94.

Williams, B.K., Nichols, J.D. and Conroy, M.J. (2002). *Analysis and Management of Animal Populations*. San Diego, California: Academic Press.

Xi, L., Watson, R. and Yip, P.S.F. (2008). The minimum capture proportion for reliable estimation in capture-recapture models. *Biometrics,* 64, 242-249.

Survey Methodology, December 2015
Vol. 41, No. 2, pp. 377-387
Statistics Canada, Catalogue No. 12-001-X

Model-assisted optimal allocation for planned domains using composite estimation

Wilford B. Molefe and Robert Graham Clark[1]

Abstract

This paper develops allocation methods for stratified sample surveys where composite small area estimators are a priority, and areas are used as strata. Longford (2006) proposed an objective criterion for this situation, based on a weighted combination of the mean squared errors of small area means and a grand mean. Here, we redefine this approach within a model-assisted framework, allowing regressor variables and a more natural interpretation of results using an intra-class correlation parameter. We also consider several uses of power allocation, and allow the placing of other constraints such as maximum relative root mean squared errors for stratum estimators. We find that a simple power allocation can perform very nearly as well as the optimal design even when the objective is to minimize Longford's (2006) criterion.

Key Words: Small area estimation; Sample design; Sample size allocation; Composite estimation; Mean squared error.

1 Introduction

Sample surveys have long been used as cost-effective means for data collection but it is also the case that general purpose surveys will often not achieve adequate precision for statistics for subpopulations of interest (often called domains or areas). Domains may be geographically based areas such as states. They may also be cross-classifications of a small geographic area and a specific demographic or social group. A domain is regarded as *small* if the domain-specific sample is not large enough to produce a direct estimator with satisfactory precision.

In this paper, we suppose that stratified sampling is used with H strata defined by the small areas, indexed by $h \in U^1$. The set of all N units in the population is denoted U and the set of H strata is denoted U^1. This effectively assumes that small areas can be identified in advance, which is not always the case (Marker 2001). Even so, the survey designer may be able to make an educated guess at areas of interest, which should still result in an improved design even if new requirements for area statistics emerge after the survey has been run. The population of N_h units in stratum h is written U_h and the sample of n_h units selected by simple random sampling without replacement (SRSWOR) from stratum h is s_h. Let Y_j be the value of the characteristic of interest for the j^{th} unit in the population. The small area population mean for stratum h is $\bar{Y}_h$ and the national mean is $\bar{Y}$. The corresponding sample estimators are $\bar{y}_h$ and $\bar{y}$, respectively; $\bar{y}_h = n_h^{-1} \sum_{j \in s_h} y_j$ and $\bar{y} = \sum_{h \in U^1} P_h \bar{y}_h$, where $P_h = N_h / N$. Let the sampling variances be $v_h = \text{var}_p(\bar{y}_h)$ and $v = \text{var}_p(\bar{y})$.

Longford (2006) considers the problem of optimal sample sizes for small area estimation for this design. The approach is based on minimizing the weighted sum of the mean squared errors of the planned small area mean estimators and an overall estimator of the mean. The weight attached to each area is proportional to the area population raised to the q^{th} power, so the value of $0 \le q \le 2$ specifies the

[1] Wilford B. Molefe, Department of Statistics, University of Botswana. E-mail: molefewb@mopipi.ub.bw; Robert Graham Clark, National Institute for Applied Statistics Research Australia, University of Wollongong. E-mail: rclark@uow.edu.au.

relative importance of larger compared to smaller areas. The mean squared error of the all-strata mean estimator is multiplied by G, where G reflects the perceived priority of this estimator. An analytical solution exists for the case where $G = 0$, but it has undesirable practical properties, and may sometimes result in zero or minimum sample sizes for some strata. When $G > 0$, Longford (2006) suggests numerical optimization.

Choudhry, Rao and Hidiroglou (2012) investigate the use of nonlinear programming (NLP) to efficiently allocate sample to strata, when there may be bounds on stratum sample sizes, and priority on overall, stratum and cross-strata domain estimators of multiple variables. The paper mostly concentrates on design-based direct area estimators, but they also consider the objective criterion of Longford (2006) for composite estimation. For the Canadian Monthly Retail Trade Survey, they show that the Longford allocation gives extremely unequal sample sizes by strata, for q equal to 0.5, 1 and 1.5. For example, when $q = 1.5$, the highest stratum coefficient of variation (CV) is 112%, and even for $q = 0.5$, the highest coefficient of variation is 24%, which was deemed too high. It is not clear whether these CV%s refer to direct or composite estimators - such high CVs would be surprising for composite estimators, as their CVs are bounded above even as the sample size tends to zero. Choudhry et al. (2012) did not investigate whether other designs such as power allocations can give low values of Longford's criteria.

The aim of this paper is to find the best allocation to strata for a linear combination of the mean squared errors of small area composite estimators and of an overall estimator of the mean, similar to Longford (2006). In Section 2 we reformulate the objective in model-assisted terms, introduce the use of regressor variables, and derive a model-assisted composite estimator. Section 3 is devoted to optimizing the design. In Subsection 3.1 we discuss direct optimization, for example by NLP. Subsection 3.2 describes power allocation with the exponent chosen to numerically minimize the objective criterion. Section 4 is a numerical study of the various methods using the Swiss canton data of Longford (2006) and Section 5 contains conclusions.

2 Composite estimation

Composite estimators for small areas are defined as convex combinations of direct (unbiased) and synthetic (biased) estimators. A simple example is the composition $(1 - \phi_h)\,\bar{y}_h + \phi_h \bar{y}$ of the sample mean $\bar{y}_h$ for the target area h and the overall sample mean $\bar{y}$ of the target variable. The coefficients ϕ_h are set with the intent to minimise its mean squared error (MSE), see for example Rao (2003, Section 4.3). The coefficients by which the MSE is minimized depend on some unknown parameters which have to be estimated.

Better results can be obtained if there are some regressors $\mathbf{x_i}$, for which domain population means are available, as well as sample data at either unit or domain level enabling Y to be regressed on $\mathbf{x}$. A synthetic estimator for domain h is then defined by $\hat{\bar{Y}}_{h(\text{syn})} = \hat{\boldsymbol{\beta}}^{\mathbf{T}} \bar{\mathbf{X}}_{\mathbf{h}}$, where $\hat{\boldsymbol{\beta}}$ is the estimated regression coefficient, and $\bar{\mathbf{X}}_{\mathbf{h}}$ is the domain population mean of the regressor variables. An efficient direct estimator which is particularly suitable when domain sizes may be small is $\bar{y}_{hr} = \bar{y}_h + \hat{\boldsymbol{\beta}}^{\mathbf{T}}(\bar{\mathbf{x}}_{\mathbf{h}} - \bar{\mathbf{X}}_{\mathbf{h}})$ (Hidiroglou and Patak 2004) where $\bar{y}_h$ and $\bar{\mathbf{x}}_{\mathbf{h}}$ are the domain h sample means of Y and X. A composite estimator can then be constructed as $\tilde{y}_h^{\mathcal{C}} = (1 - \phi_h)\,\bar{y}_{hr} + \phi_h \hat{\boldsymbol{\beta}}^{\mathbf{T}} \bar{\mathbf{X}}_{\mathbf{h}}$.

The design-based MSE of the composite estimator is given by:

$$\text{MSE}_p\left(\tilde{y}_h^{\mathcal{C}}; \bar{Y}_h\right) = (1-\phi_h)^2\, v_{hr} + \phi_h^2\left\{v_{h(\text{syn})} + B_h^2\right\} + 2\phi_h\,(1-\phi_h)\,c_h$$

where c_h is the sampling covariance of $\bar{y}_{hr}$ and $\hat{\bar{Y}}_{h(\text{syn})}, v_{hr}$ is the sampling variance of the direct estimator $\bar{y}_{hr}, v_{h(\text{syn})}$ is the sampling variance of the synthetic estimator $\hat{\bar{Y}}_{h(\text{syn})}$, and $B_h = \boldsymbol{\beta}_{\mathbf{U}}^{\mathbf{T}}\bar{\mathbf{X}}_{\mathbf{h}} - \bar{Y}_h$ is the bias of using $\hat{\bar{Y}}_{h(\text{syn})}$ to estimate $\bar{Y}_h$, with $\boldsymbol{\beta}_{\mathbf{U}}$ denoting the approximate design-based expectation of $\hat{\boldsymbol{\beta}}$. Further,

$$\text{MSE}_p\left(\tilde{y}_h^{\mathcal{C}}; \bar{Y}_h\right) \approx (1-\phi_h)^2\, v_{h(\text{syn})} + \phi_h^2 B_h^2 \tag{2.1}$$

because $c_h \ll v_h$ and $v \ll v_h$ when the number of small areas is large, under regularity conditions.

A two-level linear model ξ conditional on the values of $\mathbf{x}$ will be assumed, with uncorrelated stratum random effects u_h and unit residuals ε_i :

$$\left.\begin{aligned} Y_i &= \boldsymbol{\beta}^{\mathbf{T}}\mathbf{x}_{\mathbf{i}} + u_h + \varepsilon_i \\ E_\xi\,[u_h] = E_\xi\,[\varepsilon_i] &= 0 \\ \text{var}_\xi\,[u_h] &= \sigma_{uh}^2 \\ \text{var}_\xi\left[\varepsilon_j\right] &= \sigma_{eh}^2 \end{aligned}\right\} \tag{2.2}$$

for $h \in U^1$ and $i \in U_h$. This implies that $\text{var}_\xi\,[Y_i] = \sigma_{uh}^2 + \sigma_{eh}^2 = \sigma_h^2$ for all $i \in U$, and that the covariance $\text{cov}_\xi\left[Y_i, Y_j\right]$ equals $\rho_h\sigma_h^2$ for units $i \neq j$ in the same strata and 0 for units from different strata, where $\rho_h = \sigma_{uh}^2 \big/ \left(\sigma_{uh}^2 + \sigma_{eh}^2\right)$. For simplicity, it will be assumed that $\rho_h = \rho$ are equal for all strata.

Under model (2.1),

$$E_\xi\,[v_{hr}] = E_\xi\left[n_h^{-1} S_{hw}^2\right] = n_h^{-1}\sigma_h^2\,(1-\rho)$$

where S_{hw}^2 is the within-stratum-h sample variance of $y_i - \boldsymbol{\beta}_{\mathbf{U}}^{\mathbf{T}}\mathbf{x}_{\mathbf{i}}$; and

$$\begin{aligned} E_\xi\left[B_h^2\right] &= E_\xi\left[\left(\bar{Y}_h - \bar{Y}_{h(\text{syn})}\right)^2\right] \approx E_\xi\left[\left(\bar{Y}_h - \boldsymbol{\beta}^{\mathbf{T}}\bar{\mathbf{X}}_{\mathbf{h}}\right)^2\right] \\ &= \text{var}_\xi\,[\bar{Y}_h] = \sigma_h^2 N_h^{-1}\left[1 + (N_h - 1)\,\rho\right]. \end{aligned}$$

To simplify expressions, we assume that n, N_h and H are all large, although we do not derive rigorous asymptotic results. Assuming that N_h is large, we firstly obtain $E_\xi\left[B_h^2\right] \approx \sigma_h^2\rho$. Substituting for $E_\xi\,[v_{hr}]$ and $E_\xi\left[B_h^2\right]$ into (2.1) we get the anticipated MSE or approximate model assisted mean squared error, denoted AMSE_h :

$$\text{AMSE}_h = E_\xi\text{MSE}_p\left(\tilde{y}_h^{\mathcal{C}}; \bar{Y}_h\right) \approx (1-\phi_h)^2\, n_h^{-1}\sigma_h^2\,(1-\rho) + \phi_h^2\sigma_h^2\rho. \tag{2.3}$$

Optimizing with respect to ϕ_h we immediately obtain the optimal weight ϕ_h as:

$$\phi_{h(\text{opt})} = (1-\rho)\left[1+(n_h-1)\rho\right]^{-1}. \tag{2.4}$$

We substitute the optimum weight (2.4) into (2.3) to obtain the approximate optimum anticipated MSE:

$$\begin{aligned}\text{AMSE}_h &= E_\xi \text{MSE}_p\left(\tilde{y}_h^C\left[\phi_{h(\text{opt})}\right]; \bar{Y}_h\right)\\ &\approx \left(n_h\rho\left[1+(n_h-1)\rho\right]^{-1}\right)^2 n_h^{-1}\sigma^2(1-\rho) + \left((1-\rho)\left[1+(n_h-1)\rho\right]^{-1}\right)^2\sigma^2\rho\\ &= \sigma_h^2\rho(1-\rho)\left[1+(n_h-1)\rho\right]^{-1}.\end{aligned}$$

3 Optimizing the design

3.1 Optimal design for F

One way of measuring the performance of designs for small area estimation is with a linear combination of the anticipated MSEs of the small area mean and overall mean estimators. Following Longford (2006), but using anticipated MSEs instead of design-based MSEs, we define the criterion

$$\begin{aligned}F &= \sum_{h\in U^1} N_h^q \text{AMSE}_h + \mathrm{GN}_+^{(q)} E_\xi \text{var}_p\left[\hat{\bar{Y}}_r\right]\\ &= \sum_{h\in U^1} N_h^q \text{AMSE}_h + \mathrm{GN}_+^{(q)} E_\xi \text{var}_p\left[\sum_{h\in U^1} P_h \bar{y}_{hr}\right]\\ &\approx \sum_{h\in U^1} N_h^q \text{AMSE}_h + \mathrm{GN}_+^{(q)} E_\xi \sum_{h\in U^1} P_h^2 n_h^{-1} S_{hw}^2\\ &= \sum_{h\in U^1} N_h^q \sigma_h^2\rho(1-\rho)\left[1+(n_h-1)\rho\right]^{-1} + \mathrm{GN}_+^{(q)} \sum_{h\in U^1} \sigma_h^2 P_h^2 n_h^{-1}(1-\rho)\end{aligned} \tag{3.1}$$

where the weights N_h^q reflect the inferential priorities for area h, with $0 \le q \le 2$, and $N_+^{(q)} = \sum_{h\in U^1} N_h^q$, and $\bar{y}_{hr}$ is the grand mean estimator defined in Section 2. This objective reflects the fact that surveys have many stakeholders, some of whom will be only concerned with one specific small area, while others will place priority only on national estimators. Estimators for small regions are often considered a priority, particularly if they correspond to administrative or governmental jurisdictions, although smaller areas may be assigned less priority than larger regions. The quantity G is a relative priority coefficient. Ignoring the goal of national estimation corresponds to $G = 0$ and ignoring the goal of small area estimation corresponds to large values of G, since when G is very large the second component in (3.1) dominates. The factor $N_+^{(q)}$ is introduced to appropriately scale for the effect of the absolute sizes of N_h^q and the number of areas on the relative priority G. The criterion in (3.1) is algebraically similar to the criterion in Longford (2006). Here, however, we adopt the model-assisted approach which treats the design-based inference as the real goal of survey sampling, but employs models to choose between valid randomization-based alternatives (e.g., Chapter 6 of Särndal, Swensson and Wretman 1992).

Suppose that national estimation has no priority $(G = 0)$, and the aim is to minimize (3.1) subject to a fixed total sampling cost function $C_f = \sum_h C_h n_h$, where C_h is the unit cost of surveying a unit in stratum h. The unique stationary point for this optimization is

$$n_{h,\text{opt.}} = \frac{C_f \sqrt{N_h^q \sigma_h^2 C_h^{-1}}}{\sum_{h \in U^1} \sqrt{N_h^q \sigma_h^2 C_h}} + \frac{1-\rho}{\rho}\left(\frac{\bar{C}\sqrt{N_h^q \sigma_h^2 C_h^{-1}}}{H^{-1}\sum_{h \in U^1} \sqrt{N_h^q \sigma_h^2 C_h}} - 1\right) \tag{3.2}$$

where $\bar{C} = H^{-1}\sum_h C_h$. We will concentrate on the case when unit costs are equal across strata, so that the constraint becomes $n = \sum_h n_h$ and (3.2) simplifies to

$$n_{h,\text{opt.}} = \frac{n\sqrt{\sigma_h^2 N_h^q}}{\sum_{h \in U^1} \sqrt{\sigma_h^2 N_h^q}} + \frac{1-\rho}{\rho}\left(\frac{\sqrt{\sigma_h^2 N_h^q}}{H^{-1}\sum_{h \in U^1} \sqrt{\sigma_h^2 N_h^q}} - 1\right). \tag{3.3}$$

If there are other active constraints (e.g., minimum stratum sample sizes or maximum stratum MSEs), or if $G > 0,$ then (3.2) and (3.3) do not apply and F must be minimized numerically, for example by NLP as in Choudhry et al. (2012).

In practice it would almost always be appropriate to set $0 \le q \le 2,$ with $q = 0$ corresponding to all areas being equally important regardless of size, and $q = 2$ giving much greater weight to larger areas. (The value of $q = 2$ would lead to proportional allocation if direct estimators were used rather than composite - see for example Bankier 1988.) In many cases $q = 1$ would be a sensible compromise. For example, this has been used to motivate power allocations (Bankier 1988) for master household samples in Vietnam and South Africa (Kalton, Brick and Lê 2005, paragraph 76, page 89).

The first term in (3.3) is the optimal allocation for the direct estimator and corresponds to power allocation (Bankier 1988). The second term will be positive for more populous areas (large N_h) and negative for less populous areas. Therefore, the allocation optimal for model-assisted composite estimation has more dispersed subsample sizes $n_{h,\text{opt.}}$ than the allocation that is optimal for direct estimators.

To understand the properties of the optimal allocation when $G > 0,$ and to provide a non-iterative method, Molefe (2011, Chapter 3) derived Taylor Series approximations to the optimal $n_h,$ based on small $\rho.$ However, the resulting approximation tended to result in very large negative and very large positive values of $n_{h,\text{opt.}}$ unless ρ is very small. (In practice, these would be truncated to either 0 or the population size, respectively.) Mathematically, the issue is apparently that the optimal n_h are quite nonlinear in ρ at $\rho = 0,$ so that Taylor Series approximations are only a good approximation in a small neighbourhood of $\rho = 0.$ Taylor Series based on small values of a function of both G and ρ were also considered but had similar difficulties, and so these approaches are not further discussed here.

3.2 Power allocation

Power allocations (Bankier 1988) are defined by

$$n_h = \frac{nN_h^p}{\sum_{h \in U^1} N_h^p} \tag{3.4}$$

for $h \in U^1$, where $0 \le p \le 1$. A special case is the square root allocation when $p = 1/2$. The exponent p is called the power of the allocation. Setting $p = 1$ results in proportional allocation and $p = 0$ results in equal allocation.

Bankier (1988) proposed choosing p based on perceived relative priorities. However, this was based on direct estimators being used in each stratum. We are interested in the case where composite estimation is to be used, and the objective is to obtain a low value for F in (3.1). We obtain numerically the value of p which minimizes F by one-dimensional optimization. We further consider imposing minimum stratum sample sizes, with p re-optimized accordingly. (Alternatively, maximum stratum MSE constraints could be imposed.)

4 Numerical study

We use data on the 26 cantons of Switzerland (Longford 2006); their population sizes range from 15,000 (Appenzell-Innerrhoden) to 1.23 million (Zürich). The population of Switzerland is 7.26 million. We assume that $n = 10{,}000$, $\rho = 0.025$ and $\sigma/\mu = 1$ (following Longford 2006). The last assumption only affects the magnitude of F and the relative root mean squared errors (RRMSE) but not the relativities across methods. It is satisfied if, for example, a prevalence rate of 50% is estimated. All calculations were performed in the R statistical environment (R Development Core Team 2012). Values of $q = 0, 0.5, 1, 1.5$ and 2, and values of $G = 0, 10$ and 100 were used, as in section 5.2 of Choudhry et al. (2012). The program used to produce all results is available in the appendix of Molefe and Clark (2014).

Six different allocations are evaluated in Tables 4.1-4.3. The value of F is shown for each design, relative to the value for equal allocation. Strata sample sizes were constrained in all allocations to lie between 1 and the population sizes, while still summing to n. The first design is equal allocation, then proportional allocation. The third design is the optimal design, which minimizes F in (3.1) by NLP subject to all stratum sample sizes being at least 1. The fourth design minimizes F subject to all stratum RRMSEs being 8% or less, which, from formula (3.1), is equivalent to a minimum stratum sample size of 113. For the third and fourth designs, NLP was carried out using the R package *Rsolnp* (Ghalanos and Theussl 2011). The fifth design is power allocation, where the exponent p is calculated to minimize F. The sixth design is power allocation with all stratum sample sizes constrained to be 113 or more, and with p calculated to minimize F reflecting these constraints. In both the fifth and sixth cases, p was calculated using the *optimize* function in R.

Table 4.1 shows the efficiency of the various methods when $G = 0$, where efficiency refers to the achieved values of F from formula (3.1), which is a weighted combination of MSEs of area composite estimators and an overall grand mean estimator. When $q = 0$, equal allocation is then optimal for F, and all of the allocation methods except proportional allocation return equal allocation. For larger values of q, Optimal for Composite is the most efficient, as expected. Imposing the area maximum RRMSE constraint of 8% increases F by 4% when $q = 2$, and has negligible effect (1.4% or less) for smaller q. The optimal power allocation has virtually identical efficiency to the optimal-for-composite allocation, both with and without the area RRMSE constraint. The unconstrained optimal-for-composite and power allocations are more efficient than proportional allocation when q is small, and about equally efficient for

$q \geq 1.5.$ When the area RRMSE constraint is imposed, these designs suffer a small penalty, but are still more efficient than proportional except when $q = 2.$

Table 4.1
Relative efficiency of stratified designs for $G = 0$

Design	$q = 0$	$q = 0.5$	$q = 1$	$q = 1.5$	$q = 2$
Equal allocation	1.000	1.000	1.000	1.000	1.000
Proportional allocation	2.117	1.340	0.887	0.637	0.493
Optimal for composite	1.000	0.933	0.786	0.627	0.488
Optimal for composite with constraints	1.000	0.933	0.787	0.636	0.509
Optimal power allocation	1.000	0.933	0.786	0.628	0.490
Optimal power with constraints	1.000	0.933	0.787	0.636	0.509

Table 4.2 shows relative efficiencies for $G = 10.$ As for when $G = 0,$ the optimal-for-composite and optimal power designs perform very similarly, with a similar effect of imposing the area RRMSE constraint. The major difference compared to $G = 0$ is that proportional allocation is more efficient when G is larger. The optimal designs, even with the constraint imposed, remain more efficient than proportional allocation except for $q \geq 1.5.$

Table 4.2
Relative efficiency of stratified designs for $G = 10$

Design	$q = 0$	$q = 0.5$	$q = 1$	$q = 1.5$	$q = 2$
Equal allocation	1.000	1.000	1.000	1.000	1.000
Proportional allocation	1.360	0.944	0.701	0.568	0.491
Optimal for composite	0.875	0.784	0.668	0.565	0.490
Optimal for composite with constraints	0.875	0.784	0.670	0.575	0.505
Optimal power allocation	0.905	0.791	0.668	0.565	0.490
Optimal power with constraints	0.905	0.790	0.670	0.575	0.505

Table 4.3 shows efficiencies for large $G\,(G = 100).$ Here, proportional allocation is close to the best design for all $q.$ It is about equivalent to the unconstrained optimal designs for all $q \geq 0.5,$ and more efficient than the constrained optimal designs for all $q \geq 1.$ The relative performance of the four optimal designs is about the same as for $G = 0$ and $G = 10.$

Table 4.3
Relative efficiency of stratified designs for $G = 100$

Design	$q = 0$	$q = 0.5$	$q = 1$	$q = 1.5$	$q = 2$
Equal allocation	1.000	1.000	1.000	1.000	1.000
Proportional allocation	0.656	0.576	0.529	0.503	0.488
Optimal for composite	0.608	0.565	0.527	0.503	0.488
Optimal for composite with constraints	0.608	0.567	0.536	0.515	0.501
Optimal power allocation	0.624	0.567	0.528	0.503	0.488
Optimal power with constraints	0.612	0.568	0.536	0.515	0.501

Figure 4.1 shows the distribution of the area RRMSEs across the 26 cantons for $q \in \{0.5, 1, 1.5, 2\}$ when $G = 0$ for the four optimal designs. The results for $q = 0$ are not shown because the canton

sample sizes are then all equal for the optimal designs. The optimal for composite allocation (top left) shows a fairly tight range of area RRMSEs when $q = 0.5,$ becoming more dispersed as q increases. The maximum RRMSEs are 6.6%, 9.4%, 13.8% and 15.6% for $q = 0.5, 1, 1.5$ and 2, respectively. Thus, for $q \geq 1,$ some of the RRMSEs are undesirably large. The optimal for composite allocation with constraints forces all area RRMSEs to be 8% or less, shown by the top right panel. The bottom two panels show the corresponding optimal power allocations. The unconstrained power allocation is broadly similar to the unconstrained optimal for composite allocation, but less dispersed, with lower maximum area RRMSEs. The two constrained designs are very similar.

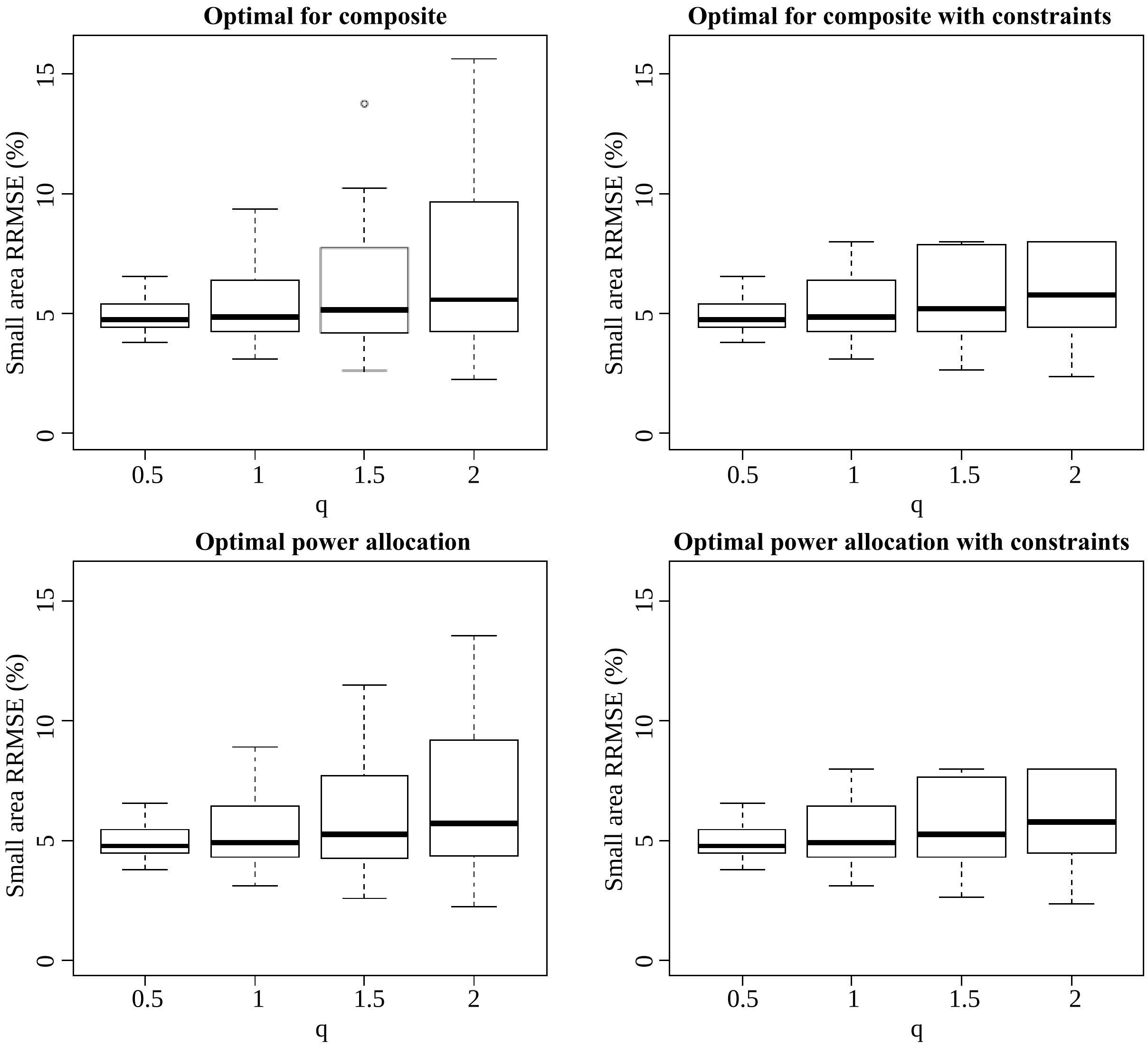


Figure 4.1 Distribution of anticipated relative root mean squared errors (RRMSE) (%) of estimated strata means for 4 allocations for various q with $G = 0$.

Table 4.4 shows the values of the optimal exponents calculated for the optimal power designs for each q and G. When G is 0 or 10, the optimal exponent p of the power allocation is very close to $q/2$, where q is the exponent in the definition of F in (3.1). For $G = 100$, the optimal exponent is quite close to 1, reflecting that for large G, F essentially reflects the variance of the grand mean, so that proportional allocation is nearly optimal. Table 4.5 shows the optimal power exponents when the area RRMSE constraints are applied. Applying these constraints has little effect on the optimal p.

Table 4.4
Optimal exponent in power allocation by G and q

	$q = 0$	$q = 0.5$	$q = 1$	$q = 1.5$	$q = 2$
$G = 0$	0.000	0.277	0.557	0.837	1.111
$G = 10$	0.293	0.500	0.721	0.912	1.050
$G = 100$	0.730	0.852	0.936	0.983	1.008

Table 4.5
Optimal exponent in power allocation by G and q with constraint on strata RRMSEs

	$q = 0$	$q = 0.5$	$q = 1$	$q = 1.5$	$q = 2$
$G = 0$	0.000	0.277	0.554	0.813	1.073
$G = 10$	0.293	0.511	0.729	0.898	1.036
$G = 100$	0.859	0.907	0.945	0.979	1.007

5 Conclusions

The anticipated MSE is a sensible objective criterion for sample design, because the particular sample which will be selected is not available in advance of the survey. Hence a criterion which averages over all possible samples is appropriate. Särndal et al. (1992, Chapter 12) base their optimal designs on the anticipated variance, which similarly averages over both model realizations and sample selection, although they consider only approximately design-unbiased estimators.

When both strata composite estimators and overall estimators are a priority, it makes sense to optimise an objective criterion which is a linear combination of the relevant anticipated MSEs. Allocations which are optimal in this sense give lower values of the objective function than either proportional or equal allocation. An optimal power allocation, $n_h \propto N_h^p$ where p is obtained numerically to minimize the objective function, is simpler and avoids the possibility of negative sample sizes which need to be truncated. Under conditions, it is very nearly as efficient as the optimal allocation. When there is no priority on national estimation $(G = 0)$, the optimal exponent turns out to be close to $p = q/2$, where q is the exponent applied to stratum population sizes in the objective criterion. This removes the need to perform an optimization. Thus, we recommend an objective criterion very similar to that of Longford (2006), but we suggest a simple power allocation with $p = q/2$ when $G = 0$, rather than the optimal allocation for F. This extends the the domain of application of power allocation to surveys using stratum composite estimators.

Rather than just relying on the overall objective criterion to appropriately balance resources across strata, it may often be desirable to also impose minimum stratum sample sizes or maximum stratum RRMSEs. These were successfully implemented using NLP. In the Swiss canton example in Section 4, an upper limit of 8% for stratum RRMSEs significantly reduced the highest RRMSE with little loss in the objective criterion. More complex constraints, for example on cross-strata domains or for multiple variables of interest, could also be implemented using NLP.

Acknowledgements

The authors wish to thank Professors Raymond Chambers and David Steel for their helpful suggestions on this paper.

Appendix

Derivation of (3.2)

The steps of this derivation are similar to Longford (2006) although F is defined differently and unequal costs are allowed. A stationary point of (3.1) subject to $C_f = \sum C_h n_h$ is given by

$$\begin{aligned} 0 &= \frac{\partial F}{\partial n_h} + \lambda C_h \\ &= -N_h^q \sigma_h^2 \rho^2 (1-\rho)(1+(n_h-1)\rho)^{-2} + \lambda C_h. \end{aligned}$$

Writing $\gamma = \lambda \rho^{-2}(1-\rho)^{-1}$ and rearranging gives

$$\begin{aligned} (1+(n_h-1)\rho)^{-2} &= \gamma C_h N_h^{-q} \sigma_h^{-2} \\ 1+(n_h-1)\rho &= \gamma^{-1/2}\sqrt{C_h^{-1} N_h^q \sigma_h^2} \qquad \text{(A.1)} \\ n_h &= \gamma^{-1/2}\rho^{-1}\sqrt{C_h^{-1} N_h^q \sigma_h^2} - \frac{1-\rho}{\rho}. \end{aligned}$$

Substituting into the constraint $C_f = \sum C_h n_h$ and solving for γ gives

$$\gamma^{-1/2} = \frac{C_f \rho + (1-\rho) H\bar{C}}{\sum_h \sqrt{\sigma_h^2 N_h^q C_h^{-1}}}$$

where $\bar{C} = H^{-1}\sum_h C_h$. Substituting back into (A.1) and rearranging gives the result.

References

Bankier, M.D. (1988). Power allocations: Determining sample sizes for subnational areas. *The American Statistician*, 42, 174-177.

Choudhry, G.H., Rao, J.N.K. and Hidiroglou, M.A. (2012). On sample allocation for efficient domain estimation. *Survey Methodology*, 38, 1, 23-29.

Ghalanos, A., and Theussl, S. (2011). Rsolnp: General Non-linear Optimization Using Augmented Lagrange Multiplier Method. http://cran.r-project.org/package=Rsolnp, R package version 1.11.

Hidiroglou, M.A., and Patak, Z. (2004). Domain estimation using linear regression. *Survey Methodology*, 30, 1, 67-78.

Kalton, G., Brick, J. and Lê, T. (2005). *Household Sample Surveys in Developing and Transition Countries*, United Nations: Statistics Division, Department of Economic and Social Affairs, no. 96 in Series F, http://unstats.un.org/unsd/hhsurveys/pdf/Household_surveys.pdf [accessed 24-May-2013].

Longford, N.T. (2006). Sample size calculation for small-area estimation. *Survey Methodology*, 32, 1, 87-96.

Marker, D.A. (2001). Producing small area estimates from National Surveys: Methods of minimizing use of indirect estimators. *Survey Methodology*, 27, 2, 183-188.

Molefe, W., and Clark, R.G. (2014). Model-assisted optimal allocation for planned domains using composite estimation. http://niasra.uow.edu.au/publications/UOW167055.html, Statistics Working Paper 08-14.

Molefe, W.B. (2011). Sample Design for Small Area Estimation. Ph.D. thesis, University of Wollongong, http://ro.uow.edu.au/theses/3495.

R Development Core Team (2012). R: A Language and Environment for Statistical Computing. http://www.R-project.org/, ISBN 3-900051-07-0.

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

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

Survey Methodology, December 2015
Vol. 41, No. 2, pp. 389-401
Statistics Canada, Catalogue No. 12-001-X

Optimum allocation for a dual-frame telephone survey

Kirk M. Wolter, Xian Tao, Robert Montgomery and Philip J. Smith[1]

Abstract

Careful design of a dual-frame random digit dial (RDD) telephone survey requires selecting from among many options that have varying impacts on cost, precision, and coverage in order to obtain the best possible implementation of the study goals. One such consideration is whether to screen cell-phone households in order to interview cell-phone only (CPO) households and exclude dual-user household, or to take all interviews obtained via the cell-phone sample. We present a framework in which to consider the tradeoffs between these two options and a method to select the optimal design. We derive and discuss the optimum allocation of sample size between the two sampling frames and explore the choice of optimum p, the mixing parameter for the dual-user domain. We illustrate our methods using the *National Immunization Survey*, sponsored by the Centers for Disease Control and Prevention.

Key Words: Dual-frame surveys; Optimum allocation; Sample design; *National Immunization Survey*.

1 Introduction

Modern random digit dial (RDD) telephone surveys in the U.S. use two samples: a sample of landlines and a sample of cell-phone lines. Wolter, Smith and Blumberg (2010) provide the statistical foundations for such dual-frame telephone surveys. The present article builds on that work and demonstrates the considerations and statistical methods for allocating the total survey resources to the two sampling frames.

Because it is less costly on a per-unit basis and has a longer history of use, the landline sample is often the larger sample and the survey interview is attempted for all respondents in this sample. The interviewing protocol for the smaller cell-phone sample is configured in one of two ways: (1) attempt to complete the survey interview for all responding persons, or (2) conduct a brief screening interview to ascertain the telephone status of the respondent, and then attempt to complete the survey interview only for respondents whose telephone status is classified as cell-phone-only (CPO) (i.e., respondents who report in the screening interview that they do not have a working landline in their household). (Within the screening approach there are variations, such as interviewing both CPO respondents and others who report that there is a landline in the household but they are not reachable through the landline.) As the size of the landline-only (LLO) population (i.e., persons who have a working landline telephone in the household but do not have access to a cell phone) declines over time (Blumberg and Luke 2010), survey statisticians may consider new designs in which the cell-phone sample is the larger sample and all respondents are interviewed, while the interviewing protocol for the smaller landline sample calls for screening or taking all respondents. Yet in this article, we focus on the prevailing circumstances in the last several years in which the cell-phone sample is typically the smaller sample and a take-all or screening protocol is used for respondents in this sample.

We shall develop the methods for optimum allocation under ideal assumptions that the sample sizes refer to completed cases (i.e., no nonresponse); that there is essentially a one-to-one relationship between the sampling units (telephone numbers) and the analytical units (e.g., households) in the landline

[1] Kirk M. Wolter, Xian Tao and Robert Montgomery, NORC at the University of Chicago, 55 East Monroe Street, Suite 3000, Chicago, IL 60603. E-mail: wolter-kirk@norc.org; Philip J. Smith, Centers for Disease Control and Prevention, National Center for Immunization and Respiratory Disease, Immunization Services Division, MS A-19, 1600 Clifton Road, NE, Atlanta, GA 30333. E-mail: pzs6@cdc.gov.

population; that there is essentially a one-to-one relationship between the sampling units and the analytical units in the cell-phone population; and that all units in the target population are included in at least one of the two sampling frames. Given these assumptions, each and every specific analytic unit is linked to a landline, a cell-phone line, or both a landline and a cell-phone line, and is linked to at most one landline and at most one cell-phone line.

Most of the previous literature on dual-frame surveys studies estimation procedures rather than the question of allocation of the sample size to the various sampling frames, including Hartley (1962, 1974); Fuller and Burmeister (1972); Skinner and Rao (1996); and Lohr and Rao (2000, 2006). Biemer (1984) and Lepkowski and Groves (1986) looked at allocation when one frame is a subset of the other frame, as might be the case with an area sample supplemented by a special list.

To begin, we establish our notation and assumptions. Let U^A be the landline population and U^B the cell-phone population. The overall population of interest is $U = U^A \cup U^B$. Some units have both a landline and a cell phone (the dual-user population), while others have only a landline (the LLO population) or only a cell phone (the CPO population), and thus the two populations overlap as follows: $U^{ab} = U^A \cap U^B, U^a = U^A - U^{ab}$, and $U^b = U^B - U^{ab}$. U^a is the LLO domain, U^b is the CPO domain, and U^{ab} is the dual-user domain. The population sizes are $N_A = \text{card}\left(U^A\right)$, $N_B = \text{card}\left(U^B\right)$, $N_{ab} = \text{card}\left(U^{ab}\right), N_a = \text{card}\left(U^a\right)$, and $N_b = \text{card}\left(U^b\right)$. We denote the proportions in the overlap (or dual-user) population by $\alpha = N_{ab} / N_A$ and $\beta = N_{ab} / N_B$.

Let s_A be a simple random sample without replacement selected from U^A, let s_B be a simple random sample without replacement selected from U^B, and let $n_A = \text{card}(s_A)$ and $n_B = \text{card}(s_B)$ be the sample sizes (i.e., completed interviews). We assume that domain membership (a, ab, b) is not known at the time of sampling.

Let Y_i be a variable of interest for the i^{th} unit in the overall population. The population domain means and variance components are denoted by $\bar{Y}_A, \bar{Y}_B, \bar{Y}_{ab}, \bar{Y}_a, \bar{Y}_b, S_A^2, S_B^2, S_{ab}^2, S_a^2$, and S_b^2. We take the goal of the survey to be the estimation of the overall population total Y.

In what follows, we derive the optimum allocation given the take-all protocol and the screening protocols in Section 2 and Section 3, respectively. Section 4 compares the two protocols in terms of efficiency and cost and attempts to provide guidance about the circumstances under which each protocol is better. The section also explores the optimum choice of a mixing parameter p, which is used to combine the estimators from the two samples $\left(s_A \cap U^{ab}\right.$ and $\left.s_B \cap U^{ab}\right)$ that represent the dual-user population. Section 5 applies the methods to the *National Immunization Survey*, a large dual-frame telephone survey sponsored by the Centers for Disease Control and Prevention (CDC). The article closes with a brief summary in Section 6.

2 Take-all protocol

In the take-all protocol, one conducts survey interviews for all units in both samples s_A and s_B. Therefore, variable data collection costs can be approximated by the model

$$C_{TA} = c_A n_A + c_B n_B, \tag{2.1}$$

where c_A is the cost per completed interview in sample s_A and c_B is the cost per completed interview in sample s_B. The expected numbers of survey interviews in the cell-phone sample are $(1-\beta)n_B$ CPO units and βn_B dual-user units.

The unbiased estimator of the population total (Hartley 1962) is given by

$$\ddot{Y} = \hat{Y}_a + p\hat{Y}_{ab} + q\hat{Y}_{ba} + \hat{Y}_b \ , \tag{2.2}$$

where p is a mixing parameter, $q = 1 - p$, $\hat{Y}_a = (N_A / n_A)\, y_a$ is an estimator of the LLO total, $\hat{Y}_{ab} = (N_A / n_A)\, y_{ab}$ is an estimator of the dual-user total derived from the landline sample, $\hat{Y}_{ba} = (N_B / n_B)\, y_{ba}$ is an estimator of the dual-user total derived from the cell-phone sample, $\hat{Y}_b = (N_B / n_B)\, y_b$ is an estimator of the CPO total, y_a is the sum of the variable of interest for the observations in s_A and in domain U^a, y_{ab} is the sum of the variable of interest for the observations in s_A and in domain U^{ab}, y_{ba} is the sum of the variable of interest for the observations in s_B and in domain U^{ab}, and y_b is the sum of the variable of interest for the observations in s_B and in domain U^b. We examine the choice of p in Section 4.

Given fixed p, we find that the variance of $\ddot{Y}$ is

$$\mathrm{Var}\{\ddot{Y}\} = N^2 \left(\frac{Q_A^2}{n_A} + \frac{Q_B^2}{n_B} \right), \tag{2.3}$$

where $W_A = N_A / N, W_B = N_B / N$,

$$Q_A^2 = W_A^2 \left\{ (1-\alpha)\, S_a^2 + \alpha p^2 S_{ab}^2 + \alpha\,(1-\alpha)\left(\bar{Y}_a - p\bar{Y}_{ab}\right)^2 \right\},$$

and

$$Q_B^2 = W_B^2 \left\{ (1-\beta)\, S_b^2 + \beta q^2 S_{ab}^2 + \beta\,(1-\beta)\left(\bar{Y}_b - q\bar{Y}_{ab}\right)^2 \right\}.$$

The classical optimum allocation of the total sample to the two sampling frames (Cochran 1977) is defined by

$$\begin{aligned} n_{A,opt} &= \frac{KQ_A}{\sqrt{c_A}} \\ n_{B,opt} &= \frac{KQ_B}{\sqrt{c_B}}, \end{aligned} \tag{2.4}$$

where K is a constant that depends upon whether the objective of the allocation is to minimize cost subject to a constraint on variance, or to minimize variance subject to a constraint on cost. The minimum variance subject to fixed cost C_{TA} is given by

$$\min\left[\mathrm{Var}\{\ddot{Y}\}\right] = \frac{\left(\sqrt{c_A}Q_A + \sqrt{c_B}Q_B\right)^2}{C_{TA}} \ , \tag{2.5}$$

while the minimum cost subject to fixed variance V_0 is

$$\min\left[C_{TA}\right] = \frac{\left(\sqrt{c_A}Q_A + \sqrt{c_B}Q_B\right)^2}{V_0}. \tag{2.6}$$

3 Screening protocol

In the screening protocol, one conducts survey interviews for all units in the landline sample s_A. One conducts screening interviews (for telephone status) for all units in the cell-phone sample s_B and then conducts the survey interviews only for the units that screen-in as CPO. Therefore, expected data collection costs arise according to the model

$$\begin{aligned} C_{SC} &= c_A n_A + c'_B \beta n_B + c''_B (1-\beta) n_B \\ &= c_A n_A + c'''_B n_B, \end{aligned} \tag{3.1}$$

where c'_B is the cost per completed screener (to ascertain telephone status) in sample s_B, c''_B is the cost per completed screener and interview in sample s_B, and $c'''_B = c'_B \beta + c''_B (1-\beta)$. In this notation, n_A is the number of survey interviews completed amongst landline respondents and n_B is the number of completed interviews (telephone screener only for non-CPO respondents, and screener plus survey interview for CPO respondents) amongst cell-phone respondents. That is, the expected total number of completed survey interviews is $n_A + (1-\beta) n_B$.

The unbiased estimator of the overall population total is

$$\hat{Y} = \hat{Y}_A + \hat{Y}_b, \tag{3.2}$$

where $\hat{Y}_A = (N_A / n_A)\, y_A$, $\hat{Y}_b = (N_B / n_B)\, y_b$, and $y_A = y_a + y_{ab}$. The variance of the estimator is

$$\operatorname{Var}\{\hat{Y}\} = N^2 \left(\frac{R_A^2}{n_A} + \frac{R_B^2}{n_B} \right), \tag{3.3}$$

where

$$R_A^2 = W_A^2 S_A^2$$

and

$$R_B^2 = W_B^2 S_b^2 \left\{ 1 - \beta + \beta (1-\beta) \frac{\bar{Y}_b^2}{S_b^2} \right\}.$$

The optimal allocation of the total sample is

$$\begin{aligned} n_{A,\,opt} &= LR_A / \sqrt{c_A} \\ n_{B,\,opt} &= LR_B / \sqrt{c'''_B}, \end{aligned}$$

where L is a constant that depends on the fixed constraint: cost or variance. The minimum variance subject to fixed cost is given by

$$\min\left[\operatorname{Var}\{\hat{Y}\}\right] = \frac{\left(\sqrt{c_A}R_A + \sqrt{c'''_B}R_B\right)^2}{C_{SC}}, \tag{3.4}$$

and the minimum cost subject to fixed variance is

$$\min\left[C_{SC}\right] = \frac{\left(\sqrt{c_A}R_A + \sqrt{c_B'''}R_B\right)^2}{V_0}. \tag{3.5}$$

4 Comparing the take-all and screening protocols

We compare the take-all and screening protocols to establish which is the less costly or more efficient. Such a comparison can provide practical guidance to planners of future dual-frame telephone surveys.

4.1 Comparing the minimum variances and costs

Given either fixed cost or fixed variance, efficiency can be assessed in terms of the ratio

$$E = \frac{\min\left[\operatorname{Var}\left\{\hat{Y}\right\}\right]}{\min\left[\operatorname{Var}\left\{\ddot{Y}\right\}\right]} = \frac{\min\left[C_{SC}\right]}{\min\left[C_{TA}\right]} = \frac{\left(\sqrt{c_A}R_A + \sqrt{c_B'''}R_B\right)^2}{\left(\sqrt{c_A}Q_A + \sqrt{c_B}Q_B\right)^2}. \tag{4.1}$$

Values less than 1.0 favor the screening approach while values greater than 1.0 favor the take-all approach.

We will illustrate efficiency using six scenarios regarding a survey of a hypothetical adult population. For all scenarios, the population size is taken from the March 2010 Current Population Survey (http://www.census.gov/cps/data/) and the population proportions by telephone status are obtained from the January – June 2010 National Health Interview Survey (Blumberg and Luke 2010). The values are $N_A = 83{,}451{,}980$, $N_a = 15{,}162{,}402$, $N_{ab} = 68{,}289{,}578$, $N_b = 31{,}265{,}108$, $N_B = 99{,}554{,}686$, $\alpha = 0.818$, and $\beta = 0.686$. For all scenarios, the aim of the survey is taken to be the estimation of the total number of adults with a certain attribute.

The scenario specific assumptions are set forth in the following table:

Table 4.1
Definition of six scenarios for a hypothetical adult population

Scenarios	$\bar{Y}_A$	$\bar{Y}_a$	$\bar{Y}_{ab}$	$\bar{Y}_b$	$\bar{Y}_B$
1	0.791	0.750	0.800	0.750	0.784
2	0.759	0.800	0.750	0.750	0.750
3	0.500	0.500	0.500	0.500	0.500
4	0.518	0.600	0.500	0.400	0.469
5	0.209	0.250	0.200	0.250	0.216
6	0.241	0.200	0.250	0.250	0.250

The means correspond to the proportions of adults with the attribute. Scenario 1 describes a population in which the domain means are similar, with the mean of the dual-user domain being somewhat larger than the means of the CPO and LLO populations. Scenario 2 describes a population in which the mean of the LLO domain is somewhat larger than the means of the other telephone status domains. Scenario 3 reflects a population in which the means of all telephone status domains are equal. Scenario 4 reflects a population in which the mean of the LLO domain is much larger than the mean of the CPO domain.

Scenarios 5 and 6 correspond to Scenarios 1 and 2, respectively, using means equal to one minus the corresponding means. The mean of the CPO domain declines from Scenario 1 to 6.

We selected the six scenarios to illustrate various circumstances in which the means of CPO, LLO, and dual-user domains differ. Differences can arise because younger adults, Hispanics, adults living only with unrelated adult roommates, adults renting their home, and adults living in poverty tend to be CPO (Blumberg and Luke 2013). To gain insight into the relative efficiencies of the take-all and screening designs, planners of future surveys may repeat our calculations for new scenarios specified by them and tailored to the particulars of their applications.

We will consider the six scenarios using three assumed cost structures. The cost structures are intended to illuminate various circumstances in which the per-unit cost of screening is high or low relative to the cost of the survey interview, with Cost Structures 1-3 reflecting increasing relative cost of screening. All cost components are expressed in interviewing hours:

Cost Structure 1: $c'_B = 0.05, c''_B = 2.05, c_B = 2.00$ and $c_A = 1.00$

Cost Structure 2: $c'_B = 0.20, c''_B = 2.20, c_B = 2.00$ and $c_A = 1.00$

Cost Structure 3: $c'_B = 0.50, c''_B = 2.50, c_B = 2.00$ and $c_A = 1.00.$

All reflect circumstances in which the hours per case for a cell-phone interview is about 2 times larger than the hours per case for a landline interview.

Efficiencies corresponding to the various scenarios for the first cost structure are illustrated in Figure 4.1. We have prepared similar figures for the second and third cost structures, but to conserve space we do not present them here.

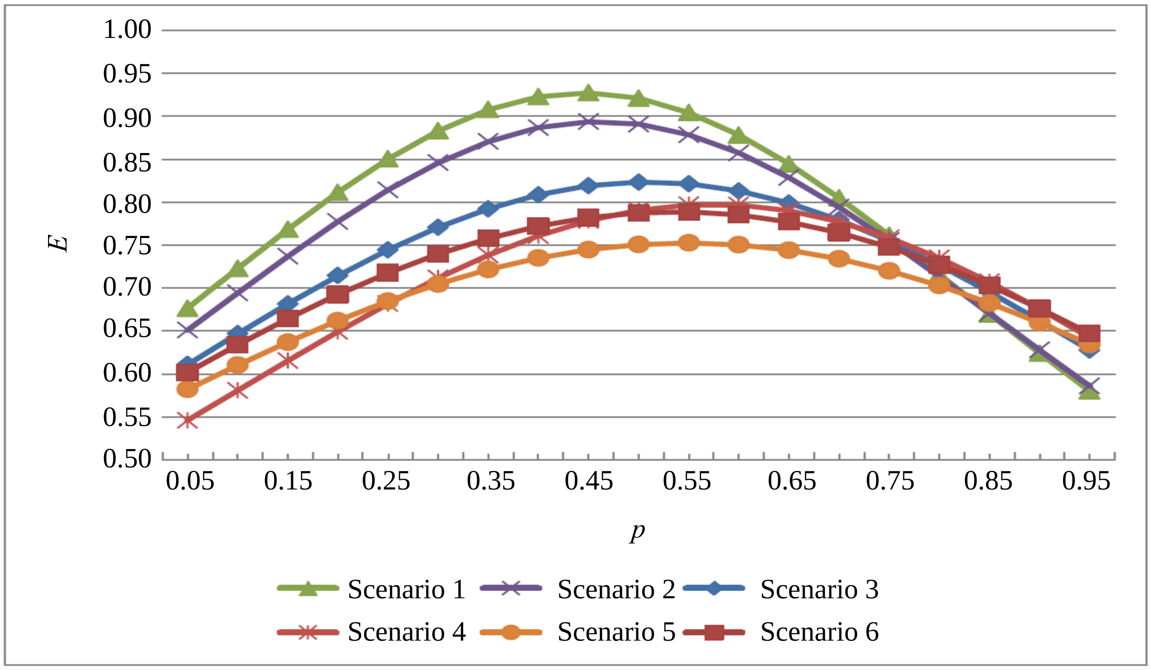


Figure 4.1 Plot of efficiency E v. mixing parameter p, given cost structure 1.

Given Cost Structure 1, the screening approach achieves the lower variance for the same fixed cost for all six scenarios. Given Cost Structure 3, in which the per-unit cost of screening is relatively much higher than in Cost Structure 1, the take-all approach achieves a smaller variance than the screening approach for half of the population scenarios. For Cost Structure 2, which entails an intermediate level of screening cost, the screening approach beats the take-all approach for all scenarios except for Scenario 1, in which the two approaches are nearly equally efficient.

The comparison between the take-all and screening protocols can be understood by examining the form of efficiency E in (4.1). The unit cost of screening is embedded only within the term $\sqrt{c_B'''}R_B$ in the numerator of $E.$ Thus, for a given scenario, the value of E must increase with increasing screening cost. For smaller screening costs, E may be less than 1.0 in which case the screening protocol will be preferred, while for larger screening costs, E may exceed 1.0 in which case the take-all protocol will be preferred.

It is also of interest to examine how the efficiency E varies with the domain means (i.e., the domain proportions), given a fixed cost structure. We see in (4.1) and in the definitions of the variance components that as long as the domain $\text{means} - \overline{Y}_b, \overline{Y}_{ab},$ and $\overline{Y}_a - \text{vary}$ reasonably together, as they do in our scenarios, the variation has relatively little or no impact on $Q_A^2, Q_B^2,$ and $R_A^2,$ and E will tend to vary more directly with $R_B^2,$ and in turn with the value of the ratio $\overline{Y}_b^2 \big/ S_b^2$ in the CPO domain. The smaller the mean in the CPO domain, the smaller this ratio will be, and in turn the smaller E will be. Thus, in each of the structures, we see smaller values of E in Scenarios 5 and 6 than in Scenarios 1 and 2, and intermediate values of E in Scenarios 3 and 4.

For the take-all protocol, the optimum p's are located at the points at which the efficiencies reach their maximum values. Table 4.2 reveals the optimum sample sizes and the optimum parameters p for each scenario and cost structure, assuming a fixed cost budget of 1,000 interviewing hours. For the screening protocol, we expect to complete $(1-\beta)n_B$ cell-phone interviews. For all population scenarios and cost structures studied here, the screening protocol obtains fewer completed cell-phone interviews than does the take-all protocol. The latter design uses resources for interviewing dual-user cases in both of the samples and requires more cell-phone interviews to provide adequate representation of CPO cases, while the former design can be more efficient about interviewing CPO cases at the price of using resources to conduct the requisite screening interviews. The optimum p's fall approximately in the range from 0.4 to 0.6 and the variance under the take-all protocol is fairly flat within this range. We examine this issue further in Section 4.2.

In summary, one may conclude from these illustrations that the screening approach is often more efficient than the take-all approach. As the cost of the screener increases relative to the cost of the interview, the outcome can tip in favor of the take-all approach. The take-all approach will be preferred for surveys in which the cost of the screener is relatively very high; otherwise, the screening protocol will be preferred. The screening approach will tend to be relatively more efficient for small values of the CPO domain mean than for large values of this mean.

Table 4.2
Sample sizes and optimum p's for the take-all and screening designs

Cost Structure	Screening Design				Take-All Design	
	n_A	n_B	$(1-\beta)n_B$	p_{opt}	n_A	n_B
			Scenario 1			
1	494	747	234	0.45	337	331
2	469	641	201	0.45	337	331
3	431	505	159	0.45	337	331
			Scenario 2			
1	506	728	229	0.45	339	330
2	481	626	197	0.45	339	330
3	443	494	155	0.45	339	330
			Scenario 3			
1	583	615	193	0.50	344	328
2	559	533	167	0.50	344	328
3	520	425	134	0.50	344	328
			Scenario 4			
1	605	582	183	0.55	377	312
2	581	506	159	0.55	377	312
3	543	405	127	0.55	377	312
			Scenario 5			
1	606	581	182	0.55	358	321
2	582	505	159	0.55	358	321
3	544	404	127	0.55	358	321
			Scenario 6			
1	618	563	177	0.55	354	323
2	594	490	154	0.55	354	323
3	557	393	123	0.55	354	323

4.2 Choosing the mixing parameter p for the take-all protocol

The optimum allocation is defined in terms of the mixing parameter, and thus it is important to consider the choice of this parameter. In the foregoing section, we saw that variance is likely not very sensitive to the choice of p within a reasonable neighborhood of optimum p. While the actual optimum p will never be known in practical applications, in this section, we describe a practical method that statisticians may use to select a reasonable, near-optimum value of p.

The landline and cell-phone samples each supply an estimator of the total in the dual-user domain, and the mixing parameter p is used to combine the two estimators into one best estimator for this domain. When the estimator of the dual-user domain derived from the landline sample is the more precise, p should be relatively large, and conversely, when the estimator from the cell-phone sample is the more precise, then $q = 1 - p$ should be relatively large. It makes good statistical sense to consider the value of p that is proportional to the expected sample size in the dual-user domain, i.e., $p_o = \alpha n_{A,\text{opt}} / \left(\alpha n_{A,\text{opt}} + \beta n_{B,\text{opt}}\right)$, where the optimum allocation is based on this choice of p. Thus, p_o is a root of the equation

$$\frac{c_A p^2}{c_B (1-p)^2} = \frac{(1-\alpha) S_a^2 + \alpha p^2 S_{ab}^2 + \alpha(1-\alpha)\left(\bar{Y}_a - p\bar{Y}_{ab}\right)^2}{(1-\beta) S_b^2 + \beta(1-p)^2 S_{ab}^2 + \beta(1-\beta)\left\{\bar{Y}_b - (1-p)\bar{Y}_{ab}\right\}^2}, \tag{4.2}$$

and, in turn, $n_{A,opt}$ and $n_{B,opt}$ are defined in terms of p_o.

From (4.2) it is apparent that p_o is a function of the $y-$variable of interest. Use of this p_o in actual practice could imply a different sample size and set of survey weights for each variable of interest, which would be unworkable. To provide a practicable solution, one might consider use of the p_o that corresponds to the survey variable $y \equiv 1$ (the population total corresponding to this variable is simply the total number of unique units on the two sampling frames). Given this approach p_o is a root of the equation

$$\frac{c_A p^2}{c_B (1-p)^2} = \frac{\alpha(1-\alpha)(1-p)^2}{\beta(1-\beta)p^2} \quad . \tag{4.3}$$

For the cost structures considered in this section, the corresponding p_o is 0.52. In Figure 4.1, one can see that this value is very close to the exact optimum p's under the various scenarios, with little loss in efficiency. Alternatively, one could evaluate (4.2) for a small set of the most important items in the survey; choose a good compromise value of p; and then define the optimum allocation in terms of this one compromise value.

5 Example: *National Immunization Survey*

5.1 Introduction

CDC has sponsored the *National Immunization Survey* (NIS) since 1994 to monitor the vaccination status of young children age $19-35$ months. The NIS uses two phases of data collection: a dual-frame RDD telephone survey of households with age-eligible children, followed by a mail survey of the vaccination providers of these children, which obtains vaccination histories for the children for each recommended vaccine. Each such child's provider-reported number of doses is compared to the recommended number of doses to determine whether the child is up-to-date (UTD). Information about the NIS is available in Smith, Hoaglin, Battaglia, Khare and Barker (2005) and the 2011 Data User's Guide (CDC 2012).

We will discuss the NIS as it was conducted in 2011. The main interview consisted of six sections, beginning with Section S, which is a brief questionnaire module that determines whether the household has age-eligible children. The interview is then terminated for ineligible households. For eligible respondents with an available vaccination record (shotcard), Section A obtains the child(ren)'s household-reported vaccination history. For all other respondents, Section B obtains a more limited and less specific amount of information about the child(ren)'s vaccinations. Section C collects demographic characteristics of the child(ren), the mother, and the household. Section D collects the names and contact information for the child(ren)'s vaccination providers and requests parental consent to contact the providers, while Section E collects information regarding current health insurance coverage.

5.2 Optimum allocation for NIS

The NIS is designed to produce estimates at the national level and for 56 non-overlapping estimation areas, consisting of 46 whole states, 6 large urban areas, and 4 rest-of-state areas. Each of these areas is a sampling stratum in the NIS design. For each of these areas, NIS is designed to minimize the cost of the survey subject to a constraint on variance: the coefficient of variation (CV) of the estimator of the vaccination coverage rate (UTD children as a proportion of all eligible children) is to be 7.5 percent at the estimation-area level, when the true rate is 50 percent.

Given the take-all protocol, the six-part survey interview is administered to all respondents in both sample. Given the screening protocol, the survey interview is administered to all respondents in the landline sample, while in the cell-phone sample, the overall interview is now in two parts: (i) the brief screener to determine telephone status and (ii) the aforementioned six-part survey interview. Dual users are screened out of the cell-phone sample.

To illustrate the optimum allocation, we take the per-unit costs to be proportional to the following values: $c'_B = 0.06$, $c''_B = 2.03$, $c_B = 1.96$, and $c_A = 1.00$. Cell-phone interviews require roughly twice as many labor hours as landline interviews. We assume the following population proportions for age-eligible children by telephone status: $W_A = 0.59$, $W_a = 0.08$, $W_{ab} = 0.51$, $W_b = 0.41$, $W_B = 0.92$, $\alpha = 0.86$, and $\beta = 0.55$. We calculated these proportions using data from the January – June 2010 National Health Interview Survey.

To estimate a vaccination coverage rate given the take-all approach, we work with the variable

$$Y_i = \begin{cases} 1, \text{ if the } i^{\text{th}} \text{ case is an age-eligible child who is UTD} \\ 0, \text{ otherwise.} \end{cases}$$

Then, the estimated vaccination coverage rate is $\ddot{Y}/N_e$, where N_e signifies the number of age-eligible children in the population (assumed known from vital statistics and related records). In accordance with the variance constraint, we take $\bar{Y}_{ae} = \bar{Y}_{abe} = \bar{Y}_{be} = 0.5$, where the subscript e signifies the mean of the age-eligible cases within the corresponding telephone status domain. Then, $\bar{Y}_d = \bar{Y}_{de} P_{de}$ and $S_d^2 = \bar{Y}_d \left(1 - \bar{Y}_d\right)$, where $d = a, ab, b$ designates the three telephone status domains and $P_{de} = N_{de}/N_d$ signifies the age-eligibility rate within domain d. Based on NIS experience, we take $P_{ae} = 0.015$, $P_{abe} = 0.03$, and $P_{be} = 0.05$, reflecting an increasing eligibility rate across the telephone status domains; that is, young child-bearing families tend to have a cell phone and further tend to be CPO. By definition, the variance is the square of the coefficient of variation times the square of the population proportion. Thus, the variance constraint is $\text{Var}\left\{\ddot{Y}/N_e\right\} = 0.075^2 \times 0.5^2$.

To estimate a vaccination coverage rate given the screening design, we work with the variable

$$Y_i = \begin{cases} 1, & \text{if } i \in s_A, & \text{and is an age-eligible child who is UTD} \\ 0, & \text{if } i \in s_A, & \text{and is not an age-eligible child or is not UTD} \\ 1, & \text{if } i \in s_B, & \text{and is CPO and is an age-eligible child who is UTD} \\ 0, & \text{if } i \in s_B, & \text{and is not CPO or is not an age-eligible child or is not UTD.} \end{cases}$$

Given these assumptions, the values of the efficiency ratio E lie below 1.0 for all values of p and from this we conclude that the screening design may be relatively less costly than the take-all design. The optimum value of p is about 0.39. However, E is quite flat in a neighborhood of the optimum and thus values of p in this neighborhood would produce similar total cost.

Given our assumptions, the optimum allocation for the take-all protocol at the optimum p is $n_A = 3{,}069$ and $n_B = 7{,}437,$ which equates to 86 NIS interviews on behalf of age-eligible children in the landline sample and 289 interviews on behalf of age-eligible children in the cell-phone sample. For the screening protocol, the optimum allocation is $n_A = 5{,}858$ and $n_B = 8{,}432,$ which we expect to yield 164 NIS interviews on behalf of age-eligible children in the landline sample and 188 NIS interviews of CPO households on behalf of their age-eligible children. These allocations apply to a single typical estimation area. Table 5.1 displays the expected sample sizes by telephone status domain given the optimum allocations. Given the screening protocol, the cell-phone sample yields an expected 4,674 dual users, which in turn reflect an expected 140 age-eligible children (who are not to be interviewed and thus are not included in the table).

Table 5.1
Expected sample sizes by telephone status domain given optimum allocations

Sample and Telephone Status Domains	Take-All Protocol		Screening Protocol	
	Expected Sample Size	Expected Age-Eligible Cases	Expected Sample Size	Expected Age-Eligible Cases
s_A	3,069	86	5,858	164
s_B	7,437	289	8,432	188
$s_A \cap U^a$	416	6	794	12
$s_A \cap U^{ab}$	2,653	80	5,064	152
$s_B \cap U^{ab}$	4,122	124	4,674	0
$s_B \cap U^b$	3,314	166	3,758	188

We developed the optimum allocations revealed here under ideal conditions in which there is no nonresponse. To prepare a sample for actual use in the NIS (or any real survey), the allocation must be adjusted by the reciprocals of the expected survey cooperation rates and by the expected design effect due to weighting and clustering.

While the extant evidence shows that the screening protocol is slightly less costly than the take-all protocol, given that both achieve the same fixed variance constraint, the take-all protocol actually provides the NIS an ongoing platform for testing and comparing both protocols. The authors continue to monitor the achieved sample composition and to conduct other specialized studies of response and nonresponse error.

6 Summary

We investigated two designs for a dual-frame telephone survey: a take-all protocol in which every respondent in the cell-phone sample is interviewed and a screening protocol in which respondents in the cell-phone sample are screened for phone status and only CPO respondents are interviewed. For each design, we derived the optimum allocation of the overall survey resources to the two sampling frames.

We studied the allocation problem given the two traditional meanings of the word “optimum”: (1) to minimize variance subject to a constraint on data collection cost, and (2) to minimize data collection cost subject to a constraint on variance. Given fixed variance, we find that the screening approach tends to achieve lower total cost than the take-all approach when the per-unit cost of screening is low relative the unit cost of the survey interview. The take-all approach can achieve the lower total cost when the per-unit cost of screening is relatively high. Similarly, given fixed total cost, the screening protocol tends to be the more efficient approach when the per-unit cost of screening is relatively low, and the take-all protocol can be the more efficient approach as the per-unit cost of screening rises. Both the landline and cell-phone samples have the capacity to produce estimators for the dual-user domain, while only the cell-phone sample can produce estimators for the CPO domain. Thus, when screening is relatively inexpensive on a per-unit basis, then it should be used to produce the largest possible sample from the CPO domain. But when screening is relatively expensive, then it is better to avoid the screening step and invest the survey resources in a larger interview sample. These results were obtained under an assumption of simple random sampling, and they may not carry over exactly to other sampling designs.

The take-all design results in two estimators for the dual-user domain, which are combined using factors of p and $1 - p$ for the estimators from the landline and cell-phone samples, respectively. We studied the optimum choice of p and gave expressions for reasonable compromise values of p. When variance (or cost) is considered as a function of p, we found that it is fairly flat in a neighborhood of the optimum. The optimum allocation itself is a function of p and we found that the allocation is relatively insensitive to choices of p within a broad neighborhood of the optimum p.

We initiated this work before 2010 at a time when the CPO population in the U.S. was only a fifth to a quarter of the total population of households. At that time it made sense to contemplate a protocol in which the larger landline sample is interviewed in its entirety and the smaller cell-phone sample is screened for CPO status. At this writing, however, the CPO population comprises more than a third of the total population of households and it is still growing. It has become reasonable to consider a new screening protocol in which the landline sample is screened for telephone status and only LLO respondents are interviewed. The foregoing allocations and findings apply to this new protocol by symmetry.

We illustrated the optimum allocations and the two interviewing protocols using the 2011 *National Immunization Survey*. The survey is designed to minimize cost under a fixed variance constraint. The NIS results are limited to the population of children age $19 - 35$ months. Similar results may or may not obtain for a general population survey or for a survey with a different structure of per-unit costs.

Acknowledgements

The authors’ kindly acknowledge suggestions for improved readability offered by the Associate Editor and referees. Disclaimer: The findings and conclusions in this paper are those of the author(s), and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

References

Biemer, P.P. (1984). Methodology for optimal dual frame sample design. *Bureau of the Census SRD Research Report CENSUS/SRD/RR-84/07* available at www.census.gov/srd/papers/pdf/rr84-07.pdf.

Blumberg, S.J., and Luke, J.V. (2010). Wireless substitution: Early release of estimates from the National Health Interview Survey, January–June 2010. *National Center for Health Statistics*. December 2010. Available at http://www.cdc.gov/nchs/nhis/releases.htm.

Blumberg, S.J., and Luke, J.V. (2013). Wireless substitution: Early release of estimates from the National Health Interview Survey, July-December 2012. *National Center for Health Statistics*. June 2013. Available at http://www.cdc.gov/nchs/nhis/releases.htm.

CDC (2012). *National Immunization Survey: A User's Guide for the 2011 Public Use Data File*. Available at http://www.cdc.gov/nchs/nis/data_files.htm.

Cochran, W.G. (1977). *Sampling Techniques*, *3rd Edition*, New York: John Wiley & Sons, Inc.

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

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.

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.

Lohr, S.L., and Rao, J.N.K. (2000). Inference from 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.

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.

Smith, P.J., Hoaglin, D.C., Battaglia, M.P., Khare, M. and Barker, L.E. (2005). Statistical Methodology of the National Immunization Survey: 1994-2002. *National Center for Health Statistics, Vital and Health Statistics*, 2(138).

Wolter, K.M., Smith, P. and Blumberg, S.J. (2010). Statistical foundations of cell-phone surveys. *Survey Methodology*, 36, 2, 203-215.

Survey Methodology, December 2015
Vol. 41, No. 2, pp. 403-425
Statistics Canada, Catalogue No. 12-001-X

Adaptive survey designs to minimize survey mode effects – a case study on the Dutch Labor Force Survey

Melania Calinescu and Barry Schouten[1]

Abstract

Assessing the impact of mode effects on survey estimates has become a crucial research objective due to the increasing use of mixed-mode designs. Despite the advantages of a mixed-mode design, such as lower costs and increased coverage, there is sufficient evidence that mode effects may be large relative to the precision of a survey. They may lead to incomparable statistics in time or over population subgroups and they may increase bias. Adaptive survey designs offer a flexible mathematical framework to obtain an optimal balance between survey quality and costs. In this paper, we employ adaptive designs in order to minimize mode effects. We illustrate our optimization model by means of a case-study on the Dutch Labor Force Survey. We focus on item-dependent mode effects and we evaluate the impact on survey quality by comparison to a gold standard.

Key Words: Mode-specific selection bias; Mode-specific measurement bias; Survey costs; Survey quality.

1 Introduction

In this paper, we propose and demonstrate the minimization of mode effects through adaptive survey designs when a survey has a single statistic or indicator. We demonstrate this method for the Dutch Labour Force Survey (LFS), which has the unemployment rate as the key indicator.

The emergence of web as a survey mode has led to a renewed discussion about mixed-mode surveys. Market research companies quickly incorporated web in their designs, official statistics institutes are slower, but also these institutes are considering mixed-mode designs including web as one of the modes. Reasons for studying mixed-mode designs include increased costs in carrying out face-to-face surveys, decreasing coverage in telephone surveys and low participation in Web surveys (Fan and Yan 2010). As a consequence, survey organizations are gradually restructuring their single mode designs into mixed-mode designs. A large-scale project Data Collection for the Social Surveys (DCSS) was initiated within the EU statistical system in 2012 to investigate mixed-mode survey designs for the Labor Force Survey (LFS), see Blanke and Luiten (2012).

It is well-known that the survey mode impacts both non-observation survey errors (item-nonresponse, unit-nonresponse and undercoverage) as well as observation survey errors (measurement error and processing error). The overall difference between two modes is usually referred to as the mode effect. The difference between the measurement errors of two modes is termed the pure mode effect or measurement effect, while the difference in undercoverage and nonresponse is termed the selection effect, see, for example, de Leeuw (2005), Dillman, Phelps, Tortora, Swift, Kohrell, Berck and Messes (2009), Vannieuwenhuyze (2013) and Klausch, Hox and Schouten (2013b) for extensive discussions. There is evidence (Jäckle, Roberts and Lynn 2010, Schouten, van den Brakel, Buelens, van der Laan, Burger and Klausch 2013b, Dillman et al. 2009) that mode effects can be large. They may lead to incomparable statistics in time or incomparable statistics over population subgroups. Assessing, minimizing and stabilizing the impact of mode effects on survey estimates has become an important goal.

[1] Melania Calinescu, Department of Mathematics, VU University Amsterdam, De Boelelaan 1081, 1081HV Amsterdam, Netherlands. E-mail: mcmelania@gmail.com; Barry Schouten, Statistics Netherlands, PO Box 24500, 2490HA Den Haag, Netherlands. E-mail: jg.schouten@cbs.nl.

There are four options to reduce the impact of mode effects in survey design and survey estimation. Thorough questionnaire design and data collection design should prevent them and survey estimation and calibration help accounting for mode effects by weighting. Careful questionnaire design reduces measurement differences between modes. This is possible by using a unified mode design for questionnaires, see Dillman et al. (2009), or by achieving an equivalent stimulus per mode, see de Leeuw (2005). Some measurement effects are, however, intrinsic to the survey mode administration process. For example, an oral versus visual presentation or the interview pace make it hard or impossible to completely remove such effects. Furthermore, questionnaire design cannot remove selection effects, although the length, layout and content may be a common cause to both measurement and selection effects. Also the history of the questions may prevent a questionnaire to be redesigned completely per mode as the survey users or stakeholders do not want to reduce the length of a questionnaire or change the wording of survey items. In summary, some mode effects will always remain, even after a thorough questionnaire redesign. If estimates of measurement effects and selection effects are available, then they can be used to design the data collection strategy of a survey, i.e., to avoid them, or to design the estimation strategy, i.e., to adjust them in future surveys.

The design option implies that some modes or sequences of modes are not applied because they are expected to lead to large mode effects with respect to some benchmark design, i.e., a survey design that is considered to be free of mode effects. The expectation of large mode effects is ideally based on pilot studies but may also lean on experience. When the choice of mode(s) is not made uniform over the whole sample but based on characteristics of persons or households, the survey design option amounts to an adaptive survey design, see Wagner (2008) and Schouten, Calinescu and Luiten (2013a). Such characteristics may be available before data collection starts or may become available during data collection in the form of paradata (i.e., data collection process data, see Kreuter 2013), leading to static and dynamic adaptive survey designs, respectively. The avoidance of mode effects by adaptive survey designs is the focus of this paper.

The adjustment option is especially interesting when there is a strong rationale or incentive to approximate true values of a statistic, i.e., when the focus is not just on comparability but also on accuracy of statistics. A drawback of the adjustment option is that it is more costly than the design option since precise estimates of mode effects are needed such that accuracy of resulting statistics is not affected. A benefit of the adjustment option is that it is more flexible. It allows for different adjustments to different survey variables, whereas the survey design option has to make an overall choice. We refer to Vannieuwenhuyze (2013), Klausch, Hox and Schouten (2013a) and Suzer-Gurtekin (2013) for a discussion of adjustment during estimation.

Another option is to stabilize mode effects, which is a useful last resort approach. Given that mode effects are conjectured to be present after questionnaire, data collection and estimation design, they can be stabilized over time by calibration of the distribution of modes in the response to some fixed distribution of modes. If the average proportion of a mode to response differs between months, the respondents to that mode get a larger weight and respondents to other modes get a smaller weight. For a discussion of this method, see Buelens and van den Brakel (2014).

In this paper, we minimize the adjusted method effect to a benchmark mode design by stratifying the population into relevant subgroups and assigning the different subgroups to different modes or sequences of modes. The adjusted method effect of a design is the difference of the nonresponse adjusted mean of that design to the nonresponse adjusted mean of the benchmark design. The adjustment follows standard procedures, i.e., calibration of response to a population distribution. Hence, the adjusted method effect is the compound of the measurement effect between the two designs and the residual selection effect between the two designs that is not removed by the nonresponse adjustment.

Adaptive survey designs and the closely resembling responsive survey designs (Heeringa and Groves 2006, Kreuter 2013) are traditionally applied to reduce nonresponse error. As far as we know, to date, only Calinescu and Schouten (2013a) have attempted to focus adaptive survey designs on measurement error or the combination of nonresponse and measurement error. The main reasons are, first, that adaptive and responsive survey designs are still in their infancy and are not widely applied, and, second, that measurement error and measurement effects are inherently hard to measure. Many applications of adaptive survey designs involve a single survey mode in which it is plausible that measurement error is relatively stable for different design choices. When the survey mode is one of the survey design features, then it is no longer plausible to make this assumption. The survey mode is, however, the most interesting design feature in adaptive survey designs due to its large quality-cost differential.

A complication that arises when including the measurement error into adaptive survey designs is that, unlike nonresponse error, it is not the result of a simple yes-no decision. A sample unit provides a response or nonresponse whereas measurement error also has a magnitude. The magnitude of the measurement error may vary per item in the survey questionnaire. This implies that with multiple survey items or variables the choice of modes is a multidimensional decision. Calinescu and Schouten (2013a) attempt to reduce this multidimensionality by using response styles (or response latencies). When a survey has only one or a few key variables, which is in fact the case for the LFS, this complication does not exist and the focus can be directly on the main variables. This is the path that we follow in the current paper.

In this paper, we, therefore, bring two novel elements: we include method effects due to modes into adaptive survey designs and we focus on a single key variable. In our demonstration for the Dutch LFS, we consider three survey modes, namely, web, phone and face-to-face, and various sequences of these modes. In recent years, the Dutch LFS design underwent a series of changes in its transition from a full face-to-face survey to a mixed-mode survey. Extensive knowledge and historical survey data on the interaction between survey design features, the survey mode in particular, and the response process is available. We use this data to estimate the various parameters that are needed for the optimization model.

The outline of the paper is as follows. In Section 2, we formulate the multi-mode optimization problem. In Section 3, we describe an algorithm for the optimization of the mode effect problem. We present the optimization results in Section 4. In Section 5, we discuss the results of the paper. Appendix A and B provide extensions to be numerical results of Section 4.

2 The multi-mode optimization problem

In this section, we construct the multi-mode optimization problem that accounts for mode effects on a single key survey variable. Apart from the survey mode, we also consider caps on the number of calls in telephone and face-to-face as design features in the optimization. In the optimization model, we allow different design features to be assigned to different subpopulations. Hence, the optimization may lead to an adaptive survey design; it does so when the optimal allocation probabilities differ over the subpopulations. In our case, the subpopulations are built on linked administrative data. Note that they could also be built based on paradata collected during the early stages of the survey. The last component to the optimization problem is given by a set of explicit quality and cost functions. In our case, the quality functions are derived from mode differences in selection and measurement bias and from requirements on the precision of statistics. As a cost function, we use the total variable costs of the survey design. In the following paragraphs, we discuss the components of the optimization problem.

We begin with the survey design features contained in the survey strategy set $\mathcal{S}$. We consider single mode and sequential mixed-mode strategies, i.e., a strategies where nonrespondents in a mode are followed-up in another mode. A single mode would be labelled as M and a sequential mixed-mode as $M_1 \rightarrow M_2$. We consider Web, telephone and face-to-face survey as the modes of interest and abbreviate them to Web, Tel and $F2F$. Examples of single mode and sequential mixed mode are Tel and $Web \rightarrow F2F$, respectively. For interview modes, we additionally consider a cap k on the number of calls, denoted as Mk. For example, $F2F3$ denotes a single mode survey strategy that uses face-to-face with a maximum of three visits. We let $Mk+$ denote the counterpart strategy where there is no explicit cap. We do not consider concurrent mixed-mode strategies (two or more modes are offered simultaneously to sample units) in this paper. This restriction is without loss of generality. It would be straightforward to apply the methodology to any set of multi-mode strategies, including hybrid forms of sequential and concurrent mixed-mode strategies. A wide or diffuse set of strategies will, however, come at the cost of a larger number of input parameters that need to be estimated. The survey strategy set $\mathcal{S}$ explicitly includes the empty strategy, denoted by Φ, which represents the case where a population unit is not sampled, i.e., no action is taken to get a response from the unit. We let $\mathcal{S}^R = \mathcal{S} \setminus \{\Phi\}$ denote the set of real, non-empty strategies.

Population units are clustered into $\mathcal{G} = \{1, \ldots, G\}$ groups given a set of characteristics X such as age, ethnicity, that can be extracted from external sources of data or from paradata. Let $p(s, g)$ be the allocation probability of strategy s to group g, i.e., a proportion $p(s, g)$ from subpopulation g is sampled and approached through strategy s. In general, it may hold that multiple strategies have non-zero allocation probabilities, so that the subpopulation is divided over multiple strategies. Define the allocation probability $p(\Phi, g)$ as the probability that a unit from subpopulation g is not included in the sample. The ratio $p(s, g)/(1 - p(\Phi, g))$ is the probability that a unit is assigned strategy s given that it has been sampled. For example, if only the allocation probabilities to the empty strategy $p(\Phi, g)$ vary and the allocation probabilities $p(s, g), \forall s \in \mathcal{S}^R$ are equal conditional on being sampled, then the design is stratified but non-adaptive. The probabilities must satisfy

$$\begin{aligned} \sum_{s \in \mathcal{S}^R} p(s, g) + p(\Phi, g) &= 1, \ \forall g \in \mathcal{G}, \\ 0 \leq p(s, g) &\leq 1, \ \forall s \in \mathcal{S}, \ g \in \mathcal{G}. \end{aligned} \tag{2.1}$$

The allocation probabilities of survey strategies assigned to subpopulations $p(s, g)$ define the decision variables in the optimization model. More generally, and analogous to sampling designs, one could allow for dependencies between population units being sampled and/or being allocated to non-empty strategies $s \in \mathcal{S}^R$. We will not add that complexity here, but assume independence.

We now discuss the quality and cost functions. We assume that the interest lies in estimating the population means of a survey variable y. Given that we consider the survey mode as one of the design features, we view the nonresponse adjusted bias on y between the proposed design and a specified benchmark design BM as the main quality function. This bias may be viewed as the adjusted method effect with respect to BM, and it is a mix of mode-specific measurement biases and remaining mode-specific nonresponse biases after adjustment. If both the proposed design and the benchmark design are single mode, then the bias is a true (adjusted) mode effect. If one of the designs is multi-mode, then the bias represents a complex mixture of mode effects, see for instance Klausch, Hox and Schouten (2014).

Let N_g be the population size of group $g, w_g = N_g / N$ be the proportion of group g in the population of size N, and $\rho(s, g)$ be the response propensity for group g if strategy s is assigned. For a specific group, we define the adjusted method effect as the nonresponse adjusted difference between the survey estimate $\bar{y}_{s,g}$ and a benchmark estimate $\bar{y}_g^{\mathrm{BM}}$ of the population mean $\bar{Y}$, where the survey estimate $\bar{y}_{s,g}$ is obtained by allocating strategy $s \in \mathcal{S}^R$ to subpopulation $g \in \mathcal{G}$. Let $D(s, g)$ denote this difference. The adjusted method effect is expressed as

$$D(s, g) = \bar{y}_{s,g} - \bar{y}_g^{\mathrm{BM}}, \ \forall s \in \mathcal{S}^R, \ g \in \mathcal{G}. \tag{2.2}$$

For convenience, we omit the adjective “adjusted”' in the following and refer to $D(s, g)$ simply as the *method effect*.

In this paper, we seek to minimize the expected absolute overall method effect with respect to a given benchmark design BM, which is the weighted average of the method effects $D(s, g)$ per stratum and strategy to BM. The expected absolute overall method effect with respect to BM is equal to

$$\bar{D}^{\mathrm{BM}} = \left| \sum_{g \in \mathcal{G}} w_g \frac{\sum_{s \in \mathcal{S}^R} p(s, g) \rho(s, g) D(s, g)}{\sum_{s \in \mathcal{S}^R} p(s, g) \rho(s, g)} \right|. \tag{2.3}$$

This objective function represents the expected shift in the time series of the key survey statistic when a redesign is implemented from the benchmark design to the adaptive design using allocation probabilities $p(s, g)$. If a survey is new or if the benchmark design was never actually fielded, the objective function represents the bias of the adaptive survey design to the benchmark design. It is, therefore, a very useful objective function. Note that $\bar{y}_{s,g}$ is a nonresponse adjusted estimate of $\bar{Y}$, while $\rho(s, g)$ is an unweighted estimate of the group g response probability in strategy s. We implicitly assume that the nonresponse adjustment does not influence the contribution of each group and strategy to the overall response. This allows us to write the objective function as in (2.4), while performing nonresponse adjustment within the optimization framework may lead to a very complex, perhaps even unsolvable, problem. We minimize the overall method effect $\bar{D}^{\mathrm{BM}}$ by optimally assigning strategies $s \in \mathcal{S}^R$ to the groups $g \in \mathcal{G}$, i.e.,

$$\underset{p(s,g)}{\text{minimize}} \ \bar{D}^{\mathrm{BM}}. \tag{2.4}$$

Ideally, $\bar{D}^{\mathrm{BM}} = 0.$ However, achieving this situation may have serious practical issues such as requiring unlimited resources. Therefore, various practical aspects such as scarcity in resources are reflected through a number of constraints in our model. A limited budget B is available to setup and run the survey. Let $c(s, g)$ be the unit cost of applying strategy s to one unit in group $g.$ The cost constraint is formulated as follows

$$\sum_{s,g} N_g\, p(s, g)\, c(s, g) \leq B. \tag{2.5}$$

To ensure a minimal precision for the survey estimate of $\bar{Y},$ a minimum number R_g of respondents per group is required. This translates to the following constraint

$$\sum_{s \in \mathcal{S}^R} N_g\, p(s, g)\, \rho(s, g) \geq R_g,\ \forall g \in \mathcal{G}. \tag{2.6}$$

In addition to the objective function, the method effect between the proposed design and the benchmark design is also part of a constraint in the optimization problem: a constraint on comparability of population subgroups. The overall method effect as an objective function could lead to an unbalanced solution. For example, let a group g be assigned a strategy s such that the corresponding $D(s, g)$ is a large negative value and the other groups $h \in \mathcal{G} \setminus \{g\}$ receive strategies that yield positive $D(s, h)$ values. The large negative $D(s, g)$ is canceled out but group g will have a very different behavior compared to the other groups, and this complicates comparisons among groups. To prevent the occurrence of such designs, we limit the absolute difference in the method effect between two groups by the following constraint

$$\max_{g,h \in \mathcal{G}} \left\{ \frac{\sum_{s \in \mathcal{S}^R} p(s, g)\, \rho(s, g)\, D(s, g)}{\sum_{s \in \mathcal{S}^R} p(s, g)\, \rho(s, g)} - \frac{\sum_{s \in \mathcal{S}^R} p(s, h)\, \rho(s, h)\, D(s, h)}{\sum_{s \in \mathcal{S}^R} p(s, h)\, \rho(s, h)} \right\} \leq M. \tag{2.7}$$

However, when

$$\frac{\sum_{s \in \mathcal{S}^R} p(s, g)\, \rho(s, g)\, D(s, g)}{\sum_{s \in \mathcal{S}^R} p(s, g)\, \rho(s, g)} - \frac{\sum_{s \in \mathcal{S}^R} p(s, h)\, \rho(s, h)\, D(s, h)}{\sum_{s \in \mathcal{S}^R} p(s, h)\, \rho(s, h)} \leq M \tag{2.8}$$

is included in the optimization problem for each pair $(g, h) \in \mathcal{G},$ then (2.7) is automatically satisfied. For practical reasons, i.e., a depletion of the sampling frame, we also introduce a constraint on the maximum sample size $S_{\max},$ i.e.,

$$\sum_{s,g} N_g\, p(s, g) \leq S_{\max}. \tag{2.9}$$

Additionally, we require that at least one $p(s, g)$ be strictly positive,

$$\sum_{s \in \mathcal{S}^R} p(s, g) > 0, \forall g \in \mathcal{G}, \tag{2.10}$$

to avoid computational errors such as division by zero in (2.8).

Objective function (2.4) together with constraints (2.1), (2.5) – (2.10) form the multi-mode optimization problem to minimize method effects against a benchmark through adaptive survey designs. This problem is a nonconvex nonlinear problem.

3 An algorithm for solving the multi-mode optimization problem

In the previous section, we introduced the quality and cost functions and constructed a multi-mode optimization problem. The subpopulation comparability constraint, i.e., the upper limit to the maximum absolute difference between group method effects, makes the problem nonconvex and hard to solve. As a consequence, when trying to solve the multi-mode optimization problem, most general-purpose nonlinear solvers cannot do better than a local optimum. Therefore, the choice of starting points in the solvers plays an important role. As such, we propose a two-step approach. In the first step, we solve a linear programming problem (LP) that addresses the linear constraints (2.1), (2.5), (2.6) and (2.9) – (2.10). In the second step, we use the optimal solution obtained in step 1 as a starting point for a local search algorithm to solve the nonconvex nonlinear problem (NNLP).

We reformulate the optimization problem to make it computationally more tractable. Since $|f(x)| = \max\{f(x), -f(x)\},$ we can rewrite the objective function via an additional variable t and impose that $f(x) \leq t$ and $-f(x) \leq t.$ Clearly, t has to be nonnegative. The constraints themselves do not change, they are simply replaced. The multi-mode optimization problem is given in (3.2).

We can derive the LP by removing the non-linear constraints on the comparability of method effects across subpopulations and by replacing the non-linear objective function by one of the linear constraints. We choose for minimization of costs as the LP objective. The resulting LP problem formulation is given by

$$\begin{aligned}
\underset{p(s,g)}{\text{minimize}} \quad & \sum_{s,g} N_g\, p(s,g)\, c(s,g) \\
\text{subject to} \quad & \sum_{s \in \mathcal{S}^R} N_g\, p(s,g)\, \rho(s,g) \geq R_g, \ \forall g \in \mathcal{G} \\
& \sum_{s,g} N_g\, p(s,g) \leq S_{\max} \\
& 0 \leq p(s,g) \leq 1, \ \forall s \in \mathcal{S}, \ g \in \mathcal{G} \\
& \sum_{s \in \mathcal{S}} p(s,g) = 1, \ \forall g \in \mathcal{G} \\
& \sum_{s \in \mathcal{S}^R} p(s,g) > 0, \ \forall g \in \mathcal{G}.
\end{aligned} \tag{3.1}$$

$$
\begin{aligned}
\text{Minimize} \quad & t \\
\text{subject to} \quad & \sum_{s,g} \frac{w_g\, p(s,g)\, \rho(s,g)\, D(s,g)}{\sum_{s' \in \mathcal{S}^R} p(s',g)\, \rho(s',g)} \leq t \\
& -\sum_{s,g} \frac{w_g\, p(s,g)\, \rho(s,g)\, D(s,g)}{\sum_{s' \in \mathcal{S}^R} p(s',g)\, \rho(s',g)} \leq t \\
& \sum_{s,g} N_g\, p(s,g)\, c(s,g) \leq B \\
& \sum_{s \in \mathcal{S}^R} N_g\, p(s,g)\, \rho(s,g) \geq R_g, \ \forall g \in \mathcal{G} \\
& \frac{\sum_{s \in \mathcal{S}^R} p(s,g)\, \rho(s,g)\, D(s,g)}{\sum_{s \in \mathcal{S}^R} p(s,g)\, \rho(s,g)} - \frac{\sum_{s \in \mathcal{S}^R} p(s,h)\, \rho(s,h)\, D(s,h)}{\sum_{s \in \mathcal{S}^R} p(s,h)\, \rho(s,h)} \leq M \qquad (3.2) \\
& \sum_{s,g} N_g\, p(s,g) \leq S_{\max} \\
& 0 \leq p(s,g) \leq 1, \ \forall s \in \mathcal{S}, \ g \in \mathcal{G} \\
& \sum_{s \in \mathcal{S}} p(s,g) = 1, \ \forall g \in \mathcal{G} \\
& \sum_{s \in \mathcal{S}^R} p(s,g) > 0, \ \forall g \in \mathcal{G} \\
& 0 \leq t.
\end{aligned}
$$

To solve the linear problem, we use the simplex method available in R in package *boot*. Our proposed two-step algorithm thus handles (3.1) in the first step. Denote by x^*_{LP} the optimal solution obtained in the LP. In the second step, x^*_{LP} is submitted to a nonlinear optimization algorithm as a starting point in order to solve (3.2). For this step, we use nonlinear algorithms available in NLOPT (see Johnson 2013), an open-source library for nonlinear optimization that can be called from R through the *nloptr* package. The NNLP second step of the algorithm is performed only if the minimal required budget found in the LP first step is smaller than or equal to the available budget B. If the minimal budget is larger, then there is no feasible solution to the optimization problem.

Given that the performance of these algorithms is problem-dependent, we choose to combine two local search algorithms in order to increase the convergence speed. Global optimization algorithms are available in the NLOPT library but their performance for our problem was significantly worse than the selected local optimization algorithms. The two selected local search algorithms are COBYLA (Constrained Optimization by Linear Approximations), introduced by Powell (1998) (see Roy 2007 for an implementation in $\mathcal{C}$) and the Augmented Lagrangian Algorithm (AUGLAG), described in Conn, Gould and Toint (1991) and Birgin and Martinez (2008). The COBYLA method builds successive linear approximations of the objective function and constraints via a simplex of $n+1$ points (in n dimensions), and optimizes these approximations in a trust region at each step. The AUGLAG method combines the objective function and the nonlinear constraints into a single function, i.e., the objective plus a penalty for any violated constraint. The resulting function is then passed to another optimization algorithm as an unconstrained problem. If the constraints are violated by the solution of this sub-problem, then the size of the penalties is increased and the process is repeated. Eventually, the process must converge to the desired solution, if that exists.

As local optimizer for the AUGLAG method we choose MMA (Method of Moving Asymptotes, introduced in Svanberg 2002), based on its performance for our numerical experiments. The strategy behind MMA is as follows. At each point $\mathbf{x_k}$, MMA forms a local approximation, that is both convex and separable, using the gradient of $f(\mathbf{x_k})$ and the constraint functions, plus a quadratic penalty term to make the approximations conservative, e.g., upper bounds for the exact functions. Optimizing the approximation leads to a new candidate point $\mathbf{x_{k+1}}$. If the constraints are met, then the process continues from the new point $\mathbf{x_{k+1}}$, otherwise, the penalty term is increased and the process is repeated.

The reason for using two local search algorithms is that AUGLAG performs better in finding the neighborhood of the global optimum but COBYLA provides a greater accuracy in locating the optimum. Therefore, the LP optimal solution is first submitted to AUGLAG and after a number of iterations, when the improvement in the objective value is below a specified threshold, the current solution of AUGLAG is submitted to COBYLA for increased accuracy. For our case study, given the precision requirements of the obtained statistics in the survey (0.5%), the results are considered accurate enough if the obtained objective value is within 10^{-4} away from the global optimum. Any further accuracy gains are completely blurred by the sampling variation and accuracy of the input parameters themselves. The computational times can run up to a few hours. Since the optimization problem does not need to be solved during data collection, this will, however, not pose practical problems.

4 Case study: The Dutch Labor Force Survey

In this section, we discuss a case study linked to the Dutch Labor Force Survey (LFS) of the years $2010-2012$. We briefly describe the design of the LFS first. We then proceed to a description of the selected design features and the selected population subgroups. Next, we explain how we have estimated the main input parameters to the optimization problem: response propensities, telephone registration propensities, variable costs and adjusted method effects with respect to two different benchmark designs. Following the estimation, we present the main optimization results. We end with a discussion of the sensitivity of optimal designs to inaccuracy of input parameters. For full details, we refer to Calinescu and Schouten (2013b).

4.1 The Dutch LFS design and redesign in 2010 – 2012

The Dutch LFS is a monthly household survey using a rotating panel with five waves at quarterly intervals. The LFS is based on an address sample using a two-stage design in which the first stage consists of municipalities and the second consists of addresses. A stratified simple random sample is drawn based on the household age, ethnicity and registered unemployment composition. All households, to a maximum of eight, that are residents at the address are invited to participate. Within each household, all members of 15 years and older are eligible; they form the potential labor force population. The LFS contains a variety of topics, from employment status, profession and working hours to educational level, but the main survey statistic is the unemployment rate.

Up to 2010, the LFS consisted of a face-to-face first wave and telephone subsequent waves. For various reasons, costs being the most important, the first wave went through a major redesign. The other waves were left unchanged, except for a few relatively small changes to the questionnaires. The redesign consisted of two phases: First, telephone was added as a survey mode, and, second, also Web was added as a survey mode. In the first phase, the face-to-face first wave was replaced by a concurrent mode design where all households with at least one listed/registered phone number were assigned to telephone and all other households to face-to-face. The listed phone numbers consist of both landline and mobile phone numbers that can be bought from commercial vendors. In the second phase, the telephone and face-to-face concurrent design was preceded by a Web invitation, resulting in a mix of a sequential and a concurrent design. All households were sent an invitation to participate through an on-line questionnaire. Nonresponding households were approached by telephone if a listed number was available and otherwise by face-to-face. The first phase was performed during 2010 and the second phase during 2012. In both years large parallel samples were drawn in order to assess method effects between the designs on the unemployment rate. The 2010 parallel run compared the old design to the intermediate concurrent design and the 2012 parallel run compared the intermediate design to the final design with all three modes.

The redesign did not change the data collection strategy per mode. In all years, the face-to-face contact strategy for the LFS first wave consists of a maximum of six visits to the address and contacts are varied over days of the week and times during the day. If no contact is made at the sixth visit, then the address is processed as a noncontact. The telephone contact strategy consists of three series of three calls. The three series are termed contact attempts and represent three different interviewer shifts. In each shift the phone number is called three times with a time lag of roughly an hour. The Web strategy is an advance letter with a login code to a website and two reminder letters with time lags of one week.

We use the 2010 – 2012 first wave LFS data to estimate various input parameters for the optimization model. In order to keep the exposition simple, and since the subsequent waves were not redesigned, we restrict ourselves to methods effects on unemployment rate estimates based on the first wave only. However, the first wave redesign may clearly have influenced the recruitment and response to waves 2 to 5. In follow-up studies at Statistics Netherlands, recruitment propensities to subsequent waves were included in the optimization problem, but we do not discuss these here. The LFS data were augmented with data from two administrative registers: the POLIS register and the UWV register. The POLIS register contains information about employments, allowances, income from employment and social benefits. The UWV register contains persons that have registered themselves as unemployed and applied for an unemployment allowance. Both registers contain relevant variables for the LFS and will be used to stratify the population.

4.2 The strategy set

The parallel runs in the LFS allow us to consider a multi-mode optimization problem with various single mode and sequential mixed-mode strategies. In the following, we abbreviate the telephone and face-to-face modes to Tel and $F2F$, respectively. Although, the sequential strategy $Web \rightarrow F2F$ is observed only for large households and for households without a registered phone, we do include this strategy in the optimization.

Since later face-to-face and telephone calls are relatively much more expensive than early calls, we also introduce a simple cap on calls. For Tel we set the cap after two calls and for $F2F$ after three calls. These values are motivated by historical survey data, e.g., after these numbers of calls the cost per call increases quickly. We let $Tel2$ and $F2F3$ denote the strategies where a cap is placed on the number of calls. $Tel2+$ and $F2F3+$ represent strategies where there is no cap and the regular contact strategy is applied. We do realize that placing a cap is not the same as restricting the number of calls in practice. This holds especially for face-to-face. With fewer calls, interviewers or interviewer staff may change behaviour and spread calls differently. At Statistics Netherlands the $Tel2$ and $F2F3$ strategies are viewed as censored strategies with shorter data collection periods, e.g., two weeks instead of four weeks. Hence, cases are removed from the interviewer workloads after the pre-specified data collection period. From this perspective, it is more reasonable to assume that the optimal contact strategy during the first two weeks of a $F2F3+$ strategy is not so different from the optimal contact strategy in $F2F3$. Still, we may expect that realized response propensities and costs in strategies with a cap are different from their simulated propensities and costs. The strategy set now becomes

$$\begin{aligned}\mathcal{S} = \; & \{Web, Tel2, Tel2+, F2F3, F2F3+, Web \rightarrow Tel2, \\ & Web \rightarrow Tel2+, Web \rightarrow F2F3, Web \rightarrow F2F3+, \Phi\},\end{aligned} \tag{4.1}$$

where Φ denotes the nonsampling strategy.

The parallel runs for the LFS in 2010 and 2012 were large. In both years the LFS sample was doubled in size for six months. Still, estimated parameters are subject to sampling variation and in case of the $Web \rightarrow F2F$ strategies possibly also to bias. We return to this issue in Section 4.6.

4.3 Population groups

In order to stratify the population, the regular LFS weighting variables were used as a starting point: unemployment office registration, age, household size, ethnicity and registered employment. Crossing the five variables led to 48 population strata (yes or no registered unemployed in household times three age classes times two household size classes times two ethnicity classes times yes or no registered employment in household). These strata were collapsed to nine disjoint strata based on their response behavior and mode effects:

1. *Registered unemployed*: Households with at least one person registered to an unemployment office (7.5% of the population).
2. 65+ *households without employment*: Households with a maximum of three persons of 15 years and older without a registration to an unemployment office, without employment and with at least one person of 65 years or older (19.8% of population)
3. *Young household members and no employment*: Households with a maximum of three persons of 15 years and older without a registration to an unemployment office, without employment,

with all persons younger than 65 years, and with at least one person between 15 and 26 years of age (2.4% of population).

4. *Non-western without employment*: Households with a maximum of three persons of 15 years and older without a registration to an unemployment office, without employment, with all persons younger than 65 years and older than 26 years of age, and at least one person of non-western ethnicity (1.5% of population).
5. *Western without employment*: Households with a maximum of three persons of 15 years and older without a registration to an unemployment office, without employment, with all persons younger than 65 years and older than 26 years of age and all persons of western ethnicity (11.0% of population).
6. *Young household member and employment*: Households with a maximum of three persons of 15 years and older without a registration to an unemployment office, with at least one employment, with all persons younger than 65 years, and with at least one person between 15 and 26 years of age (15.6% of population).
7. *Non-western and employment*: Households with a maximum of three persons of 15 years and older without a registration to an unemployment office, with at least one employment, with all persons older than 26 years of age, and at least one person of non-western ethnicity (3.9% of population).
8. *Western and employment*: Households with a maximum of three persons of 15 years and older without a registration to an unemployment office, with at least one employment, with all persons older than 26 years of age and all persons of western ethnicity (33.5% of population).
9. *Large households*: Households with more than three persons of 15 years and older without a registration to an unemployment office (4.9% of population)

The nine population strata were given informal labels in order to aid interpretation. Note, however, that the strata 7, 8 and 9 may have household members that are 65+. Furthermore, some subgroups follow from collapsing certain strata. For instance, households with at least one employment are found by combining strata 6, 7 and 8, and households with no more than three members of 15 years and older by combining all strata from 1 to 8.

In the optimization model, the nine strata were allowed different strategies and with different strategy allocation probabilities. In addition, we added precision constraints following the regular LFS on another stratification. Minimum numbers of respondents were requested based on age, ethnicity and registered unemployment. We refer again to Calinescu and Schouten (2013b) for details about these strata and corresponding precision thresholds.

4.4 Estimation of input parameters

The input parameters to the multi-mode optimization problem are subpopulation response propensities per strategy, subgroup telephone registration propensities, subgroup costs per sample unit per strategy, and subgroup adjusted method effects per strategy. We sketch the estimation of each set of parameters in the following subsections. More details can be found in Appendix A.

There are three settings that may occur when estimating input parameters: 1) The strategy is directly observed in historical survey data, 2) the strategy is only partially observed in historical survey data, i.e., only for a subset of the sample, and 3) the strategy is not observed at all.

For the LFS case study, the first setting applies to strategies Web, $Tel2+$, $F2F3+$, $Web \rightarrow Tel2+$. The second setting applies to $Web \rightarrow F2F3+$ and the third setting applies to $Tel2$, $F2F3$, $Web \rightarrow Tel2$ and $Web \rightarrow F2F3$. Sequential mixed-mode designs with face-to-face as the follow-up mode are only observed for households without a listed phone number and fall under settings 2 or 3 depending on whether a cap is placed on the number of calls. We attempted to deal with setting 2 by modeling the input parameters based on the observed differences in parameters between $Tel2+$ and $F2F3+$. We assumed that the ratio in response propensity between $F2F3(+)$ and $Tel2+$ for households with a listed phone number can be applied to $Web \rightarrow F2F3(+)$ and $Web \rightarrow Tel2+$. Furthermore, in the estimation, we assumed that strategies involving caps on the number of calls are similar to simulated strategies, i.e., by artificially restricting strategies with the full number of calls to the specified cap. Hence, we attempted to deal with setting 3 by censoring strategies. Calinescu and Schouten (2013b) elaborate these modeling steps.

For the method effect $D(s, g)$, two benchmarks were selected $\mathrm{BM}_1 = \bar{y}_{F2F3+}$ and $\mathrm{BM}_2 = 1/3 * (\bar{y}_{Web} + \bar{y}_{Tel2+} + \bar{y}_{F2F3+})$, where $\bar{y}_{\text{mode}}$ represents the average unemployment rate estimated via the indicated survey mode. The first benchmark assumes that the average unemployment rate that is estimated via a single mode face-to-face design represents the target unemployment rate. The second benchmark assumes there is no preferred mode, hence, it assigns an equal weight to each of the three modes. The $F2F3+$ benchmark is chosen because it is the traditional mode for the LFS first wave and, hence, determines the LFS time series up to 2010. Furthermore, we believe it is the mode that provides the smallest nonresponse bias for many surveys, see, e.g., Klausch et al. (2013a). It is, however, unclear whether $F2F3+$ should also be considered the mode with the smallest measurement bias. Hence, we also introduced the second benchmark to investigate the importance of the benchmark choice.

Standard errors for the estimated input parameters were approximated using bootstrap resampling per sampling stratum, following the stratified sampling design.

4.5 Optimization results

In this section, we explore the optimal allocation and minimal method effect for various budget levels, between stratum method effect levels and sample size levels

$$B \in \{160{,}000;\ 170{,}000;\ 180{,}000\}$$

$$M \in \{1\%;\ 0.5\%;\ 0.25\%\}$$

$$S_{\max} \in \{9{,}500;\ 12{,}000;\ 15{,}000\}.$$

Appendix B presents the minimal method effects for the various levels and or the two benchmark designs, BM_1 and BM_2. For the sake of brevity, here, we highlight mostly the results for BM_1, which is the former LFS design. The actual values for the non-adaptive regular three mode LFS design are

$$B = 170{,}000 \quad M = 3.00\%$$
$$S_{\max} = 11{,}000 \quad \bar{D}^{\mathrm{BM}_1} = -0.15\%.$$

Two main conclusions can be drawn from the results. First, the adaptive design is able to decrease the absolute overall method effect with respect to both benchmarks while respecting a strict constraint on the maximal between stratum method effect and keeping the budget at the current level. The only constraint that need to be relaxed in order to reduce the overall method effect is the maximal sample size. Second, for benchmark BM_2, smaller minimal overall method effects are obtained than for BM_1, with the exception of $S_{\max} = 9{,}500$. This difference is the result of the generally smaller and more similar values of the stratum method effects $D(s, g)$. We can explore the impact of the sample size constraint by comparing the optimal allocations for $S_{\max} = 9{,}500$ and $S_{\max} = 15{,}000$. Assume thresholds are set at $B = 170{,}000$, $M = 1\%$ and BM_1. Figures 4.1 and 4.2 present the optimal allocation probabilities per stratum and strategy given that a unit is sampled. Each figure can be seen as a matrix where each row represents one of the strategies in $\mathcal{S}^R$ and each column one of the 9 strata described in Section 4.3, e.g., g_1 is the registered unemployed stratum. Each cell in the matrix, i.e., intersection of a row with a column, shows the probability of assigning the corresponding strategy to the corresponding stratum. The probabilities are depicted as bars; the larger a bar, the larger the proportion of the stratum that is allocated to the strategy. The probabilities sum up to one over the strategies, i.e., over the rows. The exact values are given in the bars in case they are 20% or larger. Figure 4.1 and 4.2 show a clear shift in allocation probabilities when the sample size is allowed to increase, e.g., stratum 6 (young household member and employment) is almost fully allocated to Web and stratum 8 (western and employment) and 9 (large households) change from sequential to face-to-face only strategies.

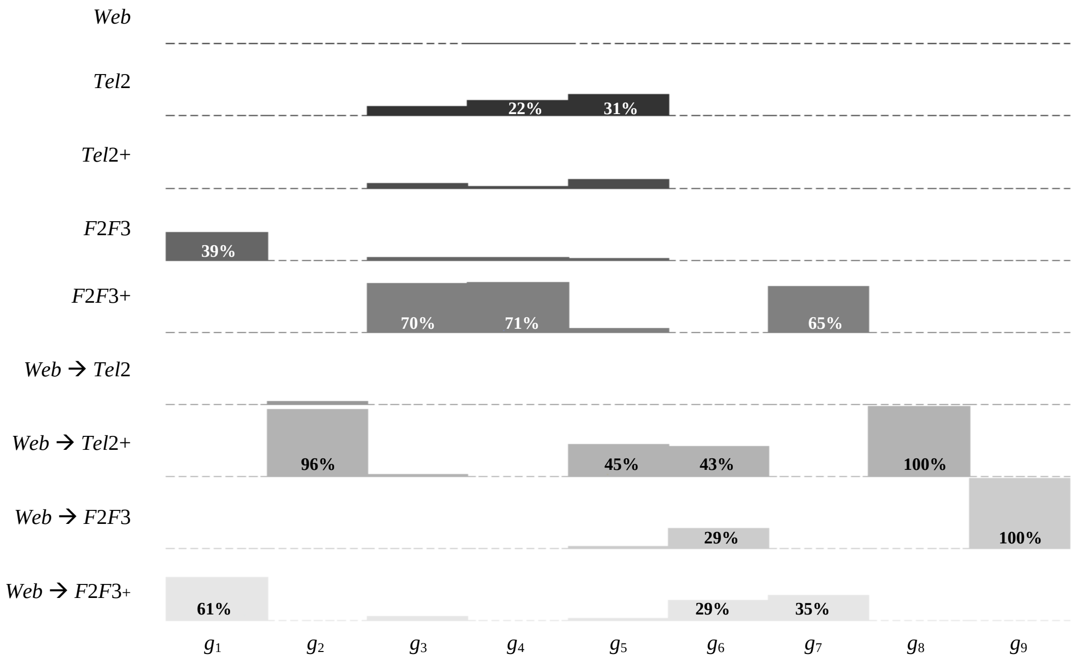


Figure 4.1 Strategy assignment given optimal solution for $S_{\max} = 9{,}500$, $B = 170{,}000$, $M = 1\%$, BM_1. The dotted line indicates that $p(s, g) = 0$.

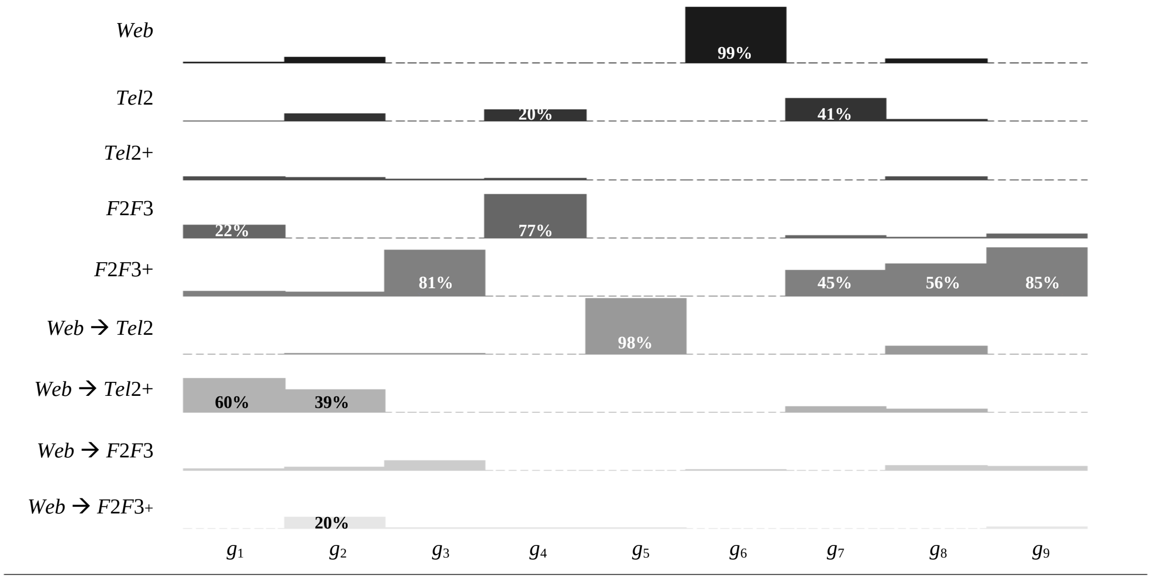


Figure 4.2 Strategy assignment given optimal solution for $S_{\max}$ = 15,000, B = 170,000, M = 1%, BM$_1$. The dotted line indicates that $p(s, g) = 0$.

The impact of the available budget can be seen very clearly for $S_{\max}$ = 12,000 and BM$_1$, where the minimal overall method effect drops from 0.10% for B = 160,000 to 0.01% for B = 180,000. The optimal allocation probabilities are shown in Figures 4.3 and 4.4. When increasing the budget, a shift takes place from telephone only strategies to a mix of face-to-face only strategies and, somewhat surprisingly, Web only strategies.

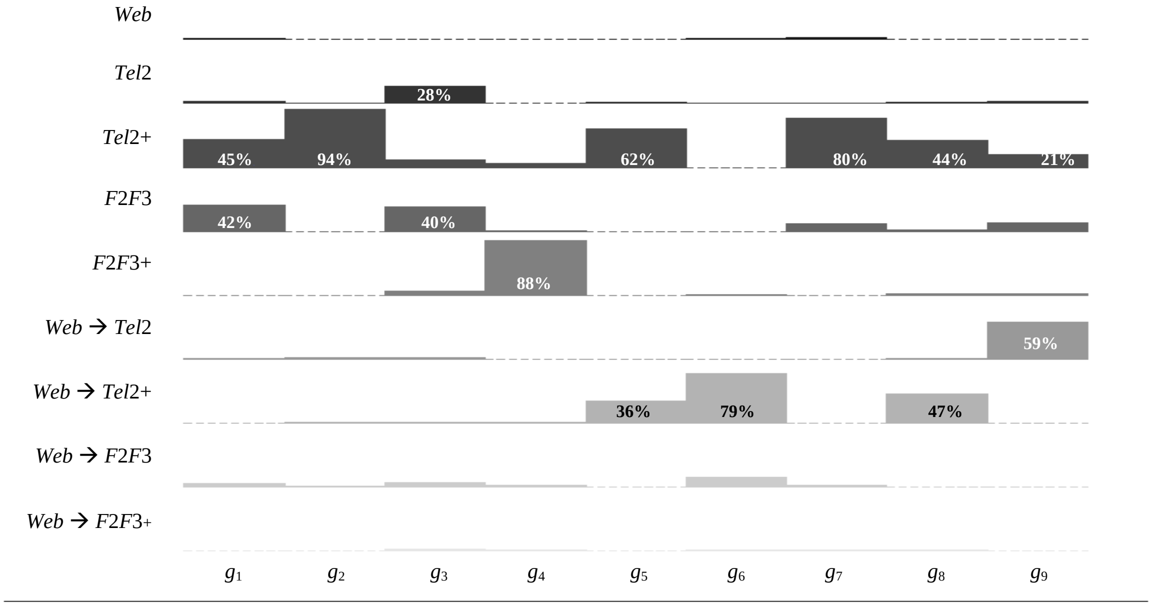


Figure 4.3 Strategy assignment given optimal solution for $S_{\max}$ = 12,000, B = 160,000, M = 1%, BM$_1$. The dotted line indicates that $p(s, g) = 0$.

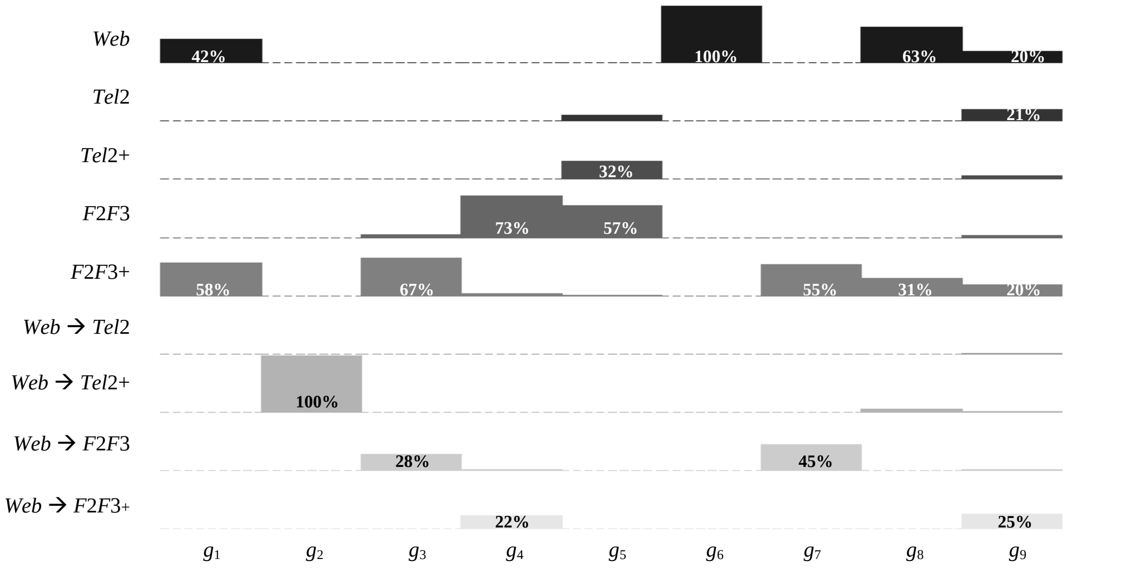


Figure 4.4 Strategy assignment given optimal solution for $S_{\max} = 12{,}000,$ $B = 180{,}000,$ $M = 1\%,$ $\mathrm{BM}_1.$ The dotted line indicates that $p(s, g) = 0.$

A range of scenarios can be investigated using a wide range of threshold values, which we leave to other papers. We conclude by mentioning that optimal allocations with many small allocation probabilities lead to very intractable data collection processes. Lower thresholds to the allocation probabilities may be added to avoid strategies that get only small numbers of cases.

4.6 Robustness of optimal designs

In this section, we briefly discuss the robustness of the optimal designs. Sensitivity analyses are beyond the scope of this paper and are part of current research.

In the estimation of the response propensities, telephone registration propensities, costs per sample unit and adjusted methods effects, we make four main assumptions; apart from assumptions about the logistic link function between response – nonresponse, telephone registration – no registration and auxiliary variables. These are:

1. Model for $Web \rightarrow F2F3$ and $Web \rightarrow F2F+$: these two strategies have only been employed for households without a listed phone number.
2. Strategies with cap on calls estimated using censoring: The strategies with a cap on calls have not been conducted and we assume that their response propensities and costs can be approximated by censoring strategies with the full contact strategy.
3. Costs linear in size allocated to strategies: We assume that costs per sample unit do not depend on the size of the sample allocated to a strategy.
4. Time stability of methods effects during 2010 – 2012: Since the parallel runs were performed in two steps, the method effects for some strategies were estimated in two steps. We implicitly assume that the methods effects for these designs have not changed over 2010 – 2012.

Furthermore, all estimated input parameters are subject to sampling variation. Consequently, we expect that certain variations in the optimal designs might occur due to inaccuracy of parameters. In order to assess robustness of optimal designs we propose two types of sensitivity analysis:

- Repeated optimization for input parameters obtained from resampled data. In other words, all historical data are resampled multiple times and for each draw an optimization is performed. The resulting optimal values for quality and costs as well as the strategy composition of the optimal designs can thus be compared across the various draws.
- Performance evaluation of the optimal design on resampled data. In other words, given observed historical data, an optimization is performed. All historical data are then resampled and for each draw the optimization input parameters are recomputed. The optimal design is applied to each set of input parameters and the corresponding quality and cost values are computed. Finally, the statistical properties of quality and cost values are assessed across all draws of input parameters.

Exploratory sensitivity analyses show that there is relatively large variation in the strategy composition of the optimal designs, but that optimal method effects $\bar{D}^{\mathrm{BM}}$ are very stable. This implies that the method effect, as objective function, is a relatively smooth function.

5 Discussion

We constructed a multi-mode optimization problem that extends the framework of adaptive survey designs to mixed-mode survey (re)designs. This framework is especially useful when it is anticipated that method effects due to a change of mode design may impact the comparability and accuracy of statistics. To our best knowledge, this is the first research attempt of its kind and can be used as a basis for minimizing method effects subject to costs and other constraints.

In the optimization model, we included three quality criteria, one cost criterion and one logistical criterion. The quality criteria were the numbers of respondents in sampling strata, which acts as a surrogate for precision, the absolute adjusted overall method effect, which is the level shift caused by the design relative to the benchmark design and may be viewed as comparability in time, and the maximal absolute difference in method effects over important subpopulations, which may be viewed as comparability over population domains. The cost criterion is the total budget of the survey. The logistic criterion is the sample size, which needs to be limited in order to avoid a quick depletion of the sampling frame. The third quality criterion, the maximal absolute difference in subpopulation method effects, is nonlinear in the decision variables (the strategy allocation probabilities) and makes the optimization problem computationally complex. Although this criterion complicates the problem, it is a useful constraint that is often put forward by survey analysts and users. In regular redesigns, this criterion is often not considered and the Dutch LFS mixed-mode design leads to relatively large differences in method effects between subpopulations. Clearly, some of the criteria may be omitted and other quality, cost or logistical criteria may be added. In a follow-up on this research at Statistics Netherlands, various other criteria, mostly logistical, are considered.

In the optimization model, the focus was on maximizing quality, reflected by comparability in time, subject to cost constraints and other constraints on quality and logistics. The objective of the optimization may, however, be changed and each of the constraints could function as the objective. For instance, one may minimize cost subject to quality and logistical constraints. One may also take a wider approach and perform several optimizations for different budget and quality levels in order to derive an informative multidimensional view on which a decision can be based.

Our attempt must be seen as merely a first step towards adaptive mixed-mode survey designs. There are various methodological and practical issues that need to be resolved. First, our approach is suited for surveys with only a few key statistics. For each of these statistics, an optimization can be performed and a weighted decision can be made. When a survey has a wide range of statistics, such an approach is not feasible. Second, the optimization leans heavily on the accuracy of its input parameters, i.e., estimated response probabilities, registered-telephone probabilities, cost parameters and mode effects in this case. It is important to assess the sensitivity of the optimization results to the accuracy of these parameters. It may be hypothesized that the objective function is relatively smooth with respect to these parameters, however, it is still important to perform sensitivity analyses. Third, it is essential to consider the sampling variation of the realized quality and costs of the optimized design when multiple waves of a survey are conducted. Such variation may be large and downsize the value of a precise optimization. Fourth, once nonlinear criteria are added to the problem, one has to rely on advanced solvers in statistical software. Even when using such solvers, convergence to global optimum is usually not assured and one has to be satisfied with local optima. For this reason, it is important to choose a useful set of starting points, including starting points that correspond to current designs. The practical issues concern the number of population strata, the number of strategies and the coordination to other surveys. Although survey administration systems and tools may support adaptive survey designs, such designs are harder to monitor and analyze. Furthermore, the tailoring of survey modes affects the size and form of interviewer workloads; interviewers may get only a specific range of subpopulations.

An important aspect of adaptive survey designs is the use of estimates for all kinds of input parameters such as response propensities, variable costs per sample unit and method effects between designs. Such estimates may not be readily available and there may only be weak historic survey data to support estimation. There are then four options: search for similar surveys that have historic support, be modest and restrictive in the choice of design features, perform a transitional period in which pilot studies and parallel runs are conducted, and develop a framework for learning and updating of parameters. In particular, designs with *Web* as one of the modes may still lack historic support for estimation in many countries, see, e.g., Mohorko, de Leeuw and Hox (2013). We also note that input parameters may gradually change in time, so that continuous updating will be needed. However, all of this is no different from a non-adaptive survey, except that now estimates are needed for relevant subpopulations instead of the overall population alone. Finally, we note that optimized adaptive designs, like optimized non-adaptive designs, provide an average, expected quality and costs. Due to sampling variation, the realized quality and costs will vary and unforeseen events may lead to deviations. Hence, monitoring and reacting to unforeseen events remain necessary.

Future research needs to address robustness of adaptive survey designs and should investigate other quality, cost and logistical criteria. It is also important that this study is replicated in order to evaluate whether the investment in terms of additional data collection and in terms of explicit optimization is worth the effort. The ultimate goal of this research is a data collection design strategy that allows for learning and updating optimization input parameters and that supports effective and efficient cost-benefit analyses in mixed-mode (re)designs. A Bayesian approach seems most promising for this purpose.

Acknowledgements

The authors would like to thank dr. Sandjai Bhulai (VU University Amsterdam) for his constructive comments on the mathematical framework presented in the current paper. The authors also thank Boukje Janssen (CBS) and Martijn Souren (CBS) for processing the raw field data for analysis and Joep Burger (CBS) for his comments that helped improve this paper.

Appendix A

Estimates of input parameters

In Section 4.4, we explain the estimation of input parameters for strategies that are observed only partially in the parallel runs. Here, we give the estimates for the response propensities, telephone registration propensities, variable costs per sample unit and adjusted method effects. Standard errors for all parameters were estimated using bootstrap resampling.

Table A2 presents the estimated response propensities $\rho(s, g)$ from available data and their corresponding standard errors. Table A1 shows the estimated propensity for a registered phone $\lambda(g)$.

Table A1
Estimated propensities for registered phone for group $g \in \mathcal{G}$ with the corresponding standard errors given in brackets

$\mathcal{G}$	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9
$\lambda(g)$	38.1% (0.9)	76.4% (1.6)	30.2% (2.0)	22.4% (2.2)	60.0% (1.1)	38.9% (0.7)	32.0% (1.3)	53.4% (0.6)	62.4% (1.2)

Table A2
Estimated response propensities per strategy s and group g with the corresponding standard errors given in brackets

$\rho(s, g)$	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9
Web	23.2% (0.3)	23.6% (0.6)	15.5% (0.6)	10.8% (0.6)	27.9% (0.4)	27.7% (0.2)	17.5% (0.5)	36.7% (0.2)	22.4% (0.5)
$Tel2$	12.2% (0.5)	31.4% (1.1)	8.5% (0.8)	4.7% (0.8)	19.7% (0.6)	13.3% (0.4)	7.2% (0.5)	18.1% (0.4)	21.2% (0.8)
$Tel2+$	20.8% (0.6)	41.3% (1.1)	15.2% (1.0)	8.6% (1.0)	31.1% (0.7)	23.8% (0.5)	14.3% (0.7)	33.3% (0.5)	37.5% (0.9)
$F2F3$	43.5% (1.5)	53.5% (1.7)	42.2% (2.4)	34.1% (2.4)	45.1% (1.1)	45.3% (0.9)	35.9% (1.5)	46.7% (0.7)	54.6% (1.4)
$F2F3+$	52.4% (1.3)	58.3% (1.6)	51.0% (2.5)	41.2% (2.2)	51.2% (1.1)	54.9% (0.8)	46.0% (1.4)	56.8% (0.7)	61.4% (1.3)
$Web \to Tel2$	28.3% (0.4)	41.0% (0.8)	20.2% (0.7)	13.9% (0.8)	36.3% (0.4)	34.0% (0.3)	20.8% (0.5)	44.5% (0.3)	23.1% (0.5)
$Web \to Tel2+$	32.8% (0.4)	48.4% (0.7)	23.8% (0.8)	17.5% (0.9)	42.1% (0.5)	41.1% (0.3)	25.8% (0.6)	52.1% (0.3)	24.4% (0.5)
$Web \to F2F3$	46.3% (0.5)	57.7% (1.0)	38.6% (1.0)	32.7% (1.0)	50.0% (0.6)	51.0% (0.4)	39.3% (0.7)	58.9% (0.4)	50.0% (0.5)
$Web \to F2F3+$	49.8% (0.5)	58.3% (0.9)	43.4% (0.9)	36.6% (0.9)	52.6% (0.5)	54.7% (0.4)	44.3% (0.6)	62.0% (0.4)	54.2% (0.5)

For the method effect $D(s, g)$, two benchmarks were selected after consultation with practitioners, i.e., $\text{BM}_1 = \bar{y}_{F2F3+}$ and $\text{BM}_2 = 1/3 * (\bar{y}_{Web} + \bar{y}_{Tel2+} + \bar{y}_{F2F3+})$, where $\bar{y}_{\text{mode}}$ represents the average unemployment rate estimated via the indicated survey mode. Tables A3 and A4 present the estimated method effects against the two benchmarks including their standard errors.

The estimates for the variable costs per sample unit plus estimated standard errors are given in Table A5. The costs are expressed relative to the $F2F3+$ strategy, which is set at one.

Table A3
Estimated method effects against benchmark $\textbf{BM}_1 = \bar{y}_{F2F3+}$ with the corresponding standard errors given in brackets

$D^{\text{BM}_1}(s, g)$	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9
Web	1.5% (1.0)	0.0% (0.5)	-2.3% (1.5)	-4.5% (3.1)	0.9% (0.7)	-0.4% (0.4)	-2.2% (1.5)	0.6% (0.5)	-0.4% (0.6)
$Tel2$	-0.2% (0.7)	-0.1% (0.1)	-2.6% (0.9)	-6.8% (1.8)	-1.0% (0.4)	-0.9% (0.3)	-1.1% (1.1)	0.2% (0.4)	-1.3% (0.4)
$Tel2+$	-0.1% (0.7)	-0.1% (0.1)	-2.3% (0.8)	-4.9% (1.7)	-0.6% (0.4)	-1.0% (0.3)	-0.8% (1.0)	-0.2% (0.3)	-1.2% (0.4)
$F2F3$	-0.5% (0.3)	-0.1% (0.1)	0.0% (0.4)	0.7% (0.6)	-0.1% (0.1)	0.0% (0.1)	0.5% (0.3)	0.3% (0.1)	0.1% (0.1)
$F2F3+$	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)	0.0% (0.0)
$Web \rightarrow Tel2$	0.9% (1.0)	0.0% (0.4)	-2.4% (1.5)	-3.4% (3.7)	-0.1% (0.6)	-0.7% (0.5)	-4.4% (1.9)	0.9% (0.5)	-0.7% (0.6)
$Web \rightarrow Tel2+$	0.9% (0.9)	-0.1% (0.3)	-3.7% (1.4)	-1.7% (3.2)	0.5% (0.7)	-0.7% (0.4)	-3.0% (1.4)	0.6% (0.5)	-0.4% (0.6)
$Web \rightarrow F2F3$	0.7% (0.6)	0.0% (0.3)	-1.2% (0.8)	-1.6% (1.4)	0.6% (0.5)	-0.3% (0.3)	-1.0% (0.8)	0.5% (0.3)	-0.2% (0.3)
$Web \rightarrow F2F3+$	0.9% (0.6)	0.0% (0.3)	-1.2% (0.8)	-2.0% (1.4)	0.6% (0.5)	-0.3% (0.3)	-1.2% (0.8)	0.4% (0.3)	-0.2% (0.3)

Table A4
Estimated method effects against benchmark $\textbf{BM}_2 = 1/3 * \left(\bar{y}_{Web} + \bar{y}_{Tel2+} + \bar{y}_{F2F3+}\right)$ with the corresponding standard errors given in brackets

$D^{\text{BM}_2}(s, g)$	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9
Web	1.0% (0.5)	0.1% (0.3)	-0.8% (0.9)	-1.4% (1.8)	0.8% (0.4)	0.1% (0.2)	-1.2% (0.8)	0.5% (0.2)	0.1% (0.3)
$Tel2$	-0.6% (0.3)	-0.1% (0.2)	-1.0% (0.6)	-3.7% (1.4)	-1.2% (0.2)	-0.5% (0.2)	-0.1% (0.8)	0.1% (0.2)	-0.8% (0.2)
$Tel2+$	-0.6% (0.2)	-0.1% (0.2)	-0.8% (0.5)	-1.7% (1.0)	-0.7% (0.2)	-0.5% (0.1)	0.2% (0.5)	-0.3% (0.1)	-0.6% (0.2)
$F2F3$	-1.0% (0.7)	-0.1% (0.2)	1.6% (0.8)	3.8% (1.6)	-0.2% (0.4)	0.5% (0.2)	1.5% (0.8)	0.2% (0.3)	0.6% (0.3)
$F2F3+$	-0.5% (0.5)	0.0% (0.2)	1.6% (0.7)	3.1% (1.4)	-0.1% (0.4)	0.5% (0.2)	1.0% (0.7)	-0.1% (0.3)	0.5% (0.3)
$Web \rightarrow Tel2$	0.4% (0.5)	0.0% (0.3)	-0.9% (1.0)	-0.3% (2.9)	-0.2% (0.4)	-0.2% (0.3)	-3.4% (1.5)	0.7% (0.3)	-0.1% (0.4)
$Web \rightarrow Tel2+$	0.5% (0.4)	0.0% (0.2)	-2.1% (0.8)	1.5% (2.0)	0.4% (0.4)	-0.2% (0.2)	-2.0% (0.8)	0.5% (0.2)	0.1% (0.3)
$Web \rightarrow F2F3$	0.3% (0.2)	0.0% (0.1)	0.4% (0.3)	1.5% (0.6)	0.5% (0.2)	0.2% (0.1)	0.0% (0.3)	0.4% (0.1)	0.3% (0.1)
$Web \rightarrow F2F3+$	0.4% (0.1)	0.0% (0.1)	0.4% (0.3)	1.1% (0.5)	0.5% (0.2)	0.2% (0.1)	-0.2% (0.3)	0.3% (0.1)	0.3% (0.1)

Table A5
Estimated relative unit costs (in euros) per strategy s and group g with the corresponding standard errors given in brackets

$c(s, g)$	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9
Web	0.03 (0.0)	0.04 (0.0)	0.04 (0.0)	0.03 (0.0)	0.04 (0.0)	0.03 (0.0)	0.03 (0.0)	0.03 (0.0)	0.03 (0.0)
$Tel2$	0.11 (0.1)	0.15 (0.1)	0.10 (0.1)	0.09 (0.1)	0.13 (0.1)	0.11 (0.1)	0.09 (0.1)	0.12 (0.0)	0.14 (0.1)
$Tel2+$	0.13 (0.1)	0.17 (0.1)	0.11 (0.1)	0.10 (0.1)	0.15 (0.1)	0.14 (0.1)	0.11 (0.1)	0.16 (0.1)	0.20 (0.2)
$F2F3$	0.84 (0.4)	0.89 (0.5)	0.83 (0.5)	0.82 (0.8)	0.86 (0.3)	0.84 (0.2)	0.81 (0.5)	0.84 (0.2)	0.89 (0.5)
$F2F3+$	1.00 (0.6)	1.00 (0.6)	1.00 (0.7)	1.00 (1.1)	1.00 (0.4)	1.00 (0.3)	1.00 (0.6)	1.00 (0.2)	1.00 (0.5)
$Web \rightarrow Tel2$	0.08 (0.0)	0.11 (0.1)	0.09 (0.1)	0.09 (0.1)	0.09 (0.0)	0.08 (0.0)	0.08 (0.0)	0.07 (0.0)	0.07 (0.0)
$Web \rightarrow Tel2+$	0.09 (0.1)	0.12 (0.1)	0.10 (0.1)	0.10 (0.1)	0.10 (0.1)	0.09 (0.0)	0.09 (0.1)	0.08 (0.0)	0.07 (0.0)
$Web \rightarrow F2F3$	0.60 (0.3)	0.66 (0.7)	0.64 (0.6)	0.70 (0.8)	0.59 (0.4)	0.56 (0.3)	0.65 (0.5)	0.51 (0.2)	0.61 (0.4)
$Web \rightarrow F2F3+$	0.71 (0.4)	0.71 (0.7)	0.80 (0.9)	0.84 (1.2)	0.73 (0.6)	0.68 (0.4)	0.81 (0.8)	0.62 (0.3)	0.71 (0.6)

Appendix B

Overview optimization results

In Section 4.5 we illustrate our approach to solve the multi-mode optimization problem for a range of input parameters. Tables B1 and B2 give a brief overview of the optimization results.

Table B1
Overview optimization results linear programming formulation - minimize costs

Sample size $(S_{\max})$	Objective value (min costs)	Benchmark	Method effect $(\bar{D}^{\mathrm{BM}})$	Max difference in mode effects (M)	Response rate
9,500	123,748.50	BM_1 BM_2	0.16% 0.29%	2.06% 3.31%	48.0%
11,000	88,408.95	BM_1 BM_2	0.05% 0.19%	5.97% 2.98%	39.9%
12,500	82,270.72	BM_1 BM_2	0.08% 0.21%	5.97% 2.98%	36.9%
15,000	74,350.44	BM_1 BM_2	0.12% 0.25%	5.97% 2.39%	29.4%

Table B2
Overview optimization results nonlinear problem - minimize average method effect in LFS

$S_{\max}$	B	BM	M	$\bar{D}^{\mathrm{BM}}$	M	$\bar{D}^{\mathrm{BM}}$	M	$\bar{D}^{\mathrm{BM}}$
9,500	160,000	BM_1 BM_2	1%	0.155% 0.170%	0.5%	Infeasible	0.25%	Infeasible
	170,000	BM_1 BM_2	1%	0.131% 0.170%	0.5%	Infeasible	0.25%	Infeasible
	180,000	BM_1 BM_2	1%	0.100% 0.170%	0.5%	Infeasible	0.25%	Infeasible
12,000	160,000	BM_1 BM_2	1%	0.097% 0.046%	0.5%	0.119% 0.046%	0.25%	0.123% 0.046%
	170,000	BM_1 BM_2	1%	0.076% 0.036%	0.5%	0.093% 0.036%	0.25%	0.101% 0.036%
	180,000	BM_1 BM_2	1%	0.009% 0.014%	0.5%	0.058% 0.014%	0.25%	0.095% 0.014%
15,000	160,000	BM_1 BM_2	1%	0.051% 0.006%	0.5%	0.094% 0.006%	0.25%	0.112% 0.006%
	170,000	BM_1 BM_2	1%	0.020% 0.004%	0.5%	0.080% 0.004%	0.25%	0.097% 0.004%
	180,000	BM_1 BM_2	1%	0.005% 0.000%	0.5%	0.058% 0.000%	0.25%	0.095% 0.000%

References

Birgin, E.G., and Martinez, J.M. (2008). Improving ultimate convergence of an augmented lagrangian method. *Optimization Methods and Software*, 23, 177-195.

Blanke, K., and Luiten, A. (2012). ESSnet project on data collection for social survey using multi modes (dcss). Paper for the UNECE Conference of European Statistics, Oct 31 - Nov 2, Geneva, Switzerland.

Buelens, B., and van den Brakel, J. (2014). Measurement error calibration in mixed-mode surveys. Forthcoming in *Sociological Methods and Research*.

Calinescu, M., and Schouten, B. (2013a). Adaptive survey designs that account for nonresponse and measurement error. Disscussion paper, Statistics Netherlands.

Calinescu, M., and Schouten, B. (2013b). Adaptive survey designs to minimize mode effects a case study on the dutch labour force survey. Disscussion paper, Statistics Netherlands.

Conn, A.R., Gould, N.I.M. and Toint, P.L. (1991). A globally convergent augmented lagrangian algorithm for optimization with general constraints and simple bounds. *SIAM Journal on Numerical Analysis*, 28, 545-572.

de Leeuw, E.D. (2005). To mix or not to mix data collection modes in surveys. *Journal of Offcial Statistics*, 21, 233-255.

Dillman, D., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J. and Messes, B. (2009). Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (ivr) and the internet. *Social Science Research*, 38, 1-18.

Fan, W., and Yan, Z. (2010). Factors affecting response rates of the web survey: a systematic review. *Computers in Human Behavior*, 26, 132-139.

Heeringa, S., and Groves, R. (2006). Responsive design for household surveys: tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(3).

Jäckle, A., Roberts, C. and Lynn, P. (2010). Assessing the effect of data collection mode on measurement. *International Statistical Review*, 78, 3-20.

Johnson, S.G. (2013). The nlopt nonlinear-optimization package. Available online at http://ab-initio.mit.edu/nlopt.

Klausch, T., Hox, J. and Schouten, B. (2013a). Assessing the mode-dependency of sample selectivity across the survey response process. Discussion paper, Statistics Netherlands.

Klausch, T., Hox, J. and Schouten, B. (2013b). Measurement effects of survey mode on the equivalence of attitudinal rating scale questions. *Sociological Methods & Research*, 42(3), 227-263.

Klausch, L., Hox, J. and Schouten, B. (2014). Evaluating bias in sequential mixed-mode surveys against single- and hybrid-mode benchmarks the case of the crime victimization survey. Discussion paper, Statistics Netherlands.

Kreuter, F. (2013). *Improving Surveys with Process and Paradata*. John Wiley & Sons, Inc.

Mohorko, A., de Leeuw, E.D. and Hox, J. (2013). Internet coverage and coverage bias in Europe: Developments across countries over time. *Journal of Offcial Statistics*, 29, 609-622.

Powell, M.J.D. (1998). Direct search algorithms for optimization calculations. *Acta Numerica*, 7, 287-336.

Roy, J.S. (2007). Stochastic optimization - scipy project. Available online at http://js2007.free.fr/.

Schouten, B., Calinescu, M. and Luiten, A. (2013a). Optimizing quality of response through adaptive survey designs. *Survey Methodology*, 39, 1, 29-58.

Schouten, B., van den Brakel, J., Buelens, B., van der Laan, J., Burger, J. and Klausch, T. (2013b). Disentangling mode-specific selection and measurement bias in social surveys. *Social Science Research*, 42, 1555-1570.

Suzer-Gurtekin, Z. (2013). *Investigating the Bias Properties of Alternative Statistical Inference Methods in Mixed-Mode Surveys*. Ph.D. thesis, University of Michigan.

Svanberg, K. (2002). A class of globally convergent optimization methods based on conservative convex separable approximations. *SIAM Journal on Optimization*, 12, 555-573.

Vannieuwenhuyze, J. (2013). *Mixed-Mode Data Collection: Basic Concepts and Analysis of Mode Effects*. Ph.D. thesis, Katholieke Universiteit Leuven.

Wagner, J. (2008). *Adaptive Survey Design to Reduce Nonresponse Bias*. Ph.D. thesis, University of Michigan.

Vol. 41, No. 2, pp. 427-442
Statistics Canada, Catalogue No. 12-001-X

Integer programming formulations applied to optimal allocation in stratified sampling

José André de Moura Brito, Pedro Luis do Nascimento Silva,
Gustavo Silva Semaan and Nelson Maculan[1]

Abstract

The problem of optimal allocation of samples in surveys using a stratified sampling plan was first discussed by Neyman in 1934. Since then, many researchers have studied the problem of the sample allocation in multivariate surveys and several methods have been proposed. Basically, these methods are divided into two classes: The first class comprises methods that seek an allocation which minimizes survey costs while keeping the coefficients of variation of estimators of totals below specified thresholds for all survey variables of interest. The second aims to minimize a weighted average of the relative variances of the estimators of totals given a maximum overall sample size or a maximum cost. This paper proposes a new optimization approach for the sample allocation problem in multivariate surveys. This approach is based on a binary integer programming formulation. Several numerical experiments showed that the proposed approach provides efficient solutions to this problem, which improve upon a 'textbook algorithm' and can be more efficient than the algorithm by Bethel (1985, 1989).

Key Words: Stratification; Allocation; Integer programming; Multivariate survey.

1 Introduction

A large part of the statistics produced by official statistics agencies in many countries come from sample surveys. Such surveys have a well-defined survey population to be covered, including the geographic location and other eligibility criteria, use appropriate frames to guide the sample selection, and apply some well-specified sample selection procedures. The use of 'standard' probability sampling procedures enables producing estimates for the target population parameters with controlled precision while having data from typically small samples of the populations, at a fraction of the cost of corresponding censuses.

When designing the sampling strategy, the survey planner often seeks to optimize precision for the most important survey estimates given an available survey budget. Stratification is an important tool that enables exploring prior auxiliary information available for all the population units by forming groups of homogeneous units, and then sampling independently from within such groups. Thus stratification is very frequently used in a wide range of sample surveys.

Here we focus on element sampling designs (Särndal, Swensson and Wretman 1992) where the frame consists of one record per population unit, and besides identification and location information, some auxiliary information is also available for each population unit. Stratified sampling involves dividing the N units in a population U into H homogeneous groups, called strata. These groups are formed considering one (or more) stratification variable(s), and such that variance within groups is small (the stratum formation problem).

[1] José André de Moura Brito and Pedro Luis do Nascimento Silva, Escola Nacional de Ciências Estatísticas (ENCE/IBGE), R.André Cavalcanti, 106, sala 403, Centro, Rio de Janeiro/RJ. E-mail: jambrito@gmail.com and pedronsilva@gmail.com; Gustavo Silva Semaan, Instituto do Noroeste Fluminense de Educação Superior, Universidade Federal Fluminense - INFES/UFF, Av. João Jasbick, s/n, Bairro Aeroporto - Santo Antônio de Pádua - RJ - CEP 28470-000. E-mail: gustavosemaan@gmail.com; Nelson Maculan, Universidade Federal do Rio de Janeiro (COPPE/UFRJ), Endereço: Av. Horácio Macedo, 2030 - CT, Bloco H, sl. 319 - Cidade Universitária, Ilha do Fundão - Rio de Janeiro, RJ - CEP 21941-914. E-mail: nelson.maculan@gmail.com.

Given a sample size n, once the strata are defined the next problem consists of specifying how many sample units should be selected in each stratum such that the variance of a specified estimator is minimized (the optimal sample allocation problem). When interest is restricted to estimating the population total (or mean) for a single survey variable, the well-known Neyman allocation (see e.g., Cochran 1977) may be used to decide on the sample allocation. Although surveys which have a single target variable are rare, Neyman's simple allocation formula may still be useful because the allocation which is optimal for a target variable may still be reasonable for other survey variables which are positively correlated with the one used to drive the optimal allocation.

When a survey must produce estimates with specified levels of precision for a number of survey variables, and these variables are not strongly correlated, a method of sample allocation that enables producing estimates with the required precision for all the survey variables is needed. In this case, we have a problem of multivariate optimal sample allocation.

According to the literature, in such cases the allocation of the overall sample size n to the strata may seek one of the following goals:

(i) the total variable survey cost C is minimized, subject to having Coefficients of Variation (CVs) for the estimates of totals of the m survey variables below specified thresholds; or

(ii) a weighted sum of variances (or relative variances) of the estimates of totals for the m survey variables is minimized.

Note that the CV is simply the square root of the relative variance.

This paper presents a new approach based on developing and applying two binary integer programming formulations that satisfy each of these two goals, while ensuring that the resulting allocation provides the global optimum. The paper is divided as follows. Section 2 reviews some key stratified sampling concepts and definitions. Section 3 describes the new approach proposed here. Section 4 provides results for a subset of numerical experiments carried out to test the proposed approach using selected population datasets. Section 5 gives some final remarks and concludes the paper. Appendix A provides information about three populations used in the numerical experiments presented in Section 4.

2 Stratified sampling and the optimal allocation problem

In stratified sampling (Cochran 1977; Lohr 2010) a population U formed by N units is divided into H strata $U_1, U_2, \ldots, U_H$ having $N_1, N_2, \ldots, N_H$ units respectively. These strata do not overlap (2.1) and together form the entire population (2.2) such that:

$$U_h \bigcap U_k = \varnothing, \quad h \neq k \tag{2.1}$$

$$\bigcup_{h=1}^{H} U_h = U \tag{2.2}$$

$$N_1 + N_2 + \ldots + N_H = \sum_{h=1}^{H} N_h = N. \tag{2.3}$$

Once the strata are defined, and given an overall sample size $n,$ an independent sample of size n_h is selected from the N_h units in stratum U_h $(h = 1, \ldots, H)$ such that $n_{\min} \leq n_h \leq N_h \; \forall h,$ where $n_{\min}$ is the smallest possible sample size in any stratum, and $n_1 + n_2 + \ldots + n_H = \sum_{h=1}^{H} n_h = n.$

A minimum sample size per stratum of $n_{\min} = 2$ is considered here, but this value may be changed as needed to accommodate specific survey requirements. A minimum sample size of one per stratum is not recommended because this might lead to solutions that require using approximate methods for variance estimation whenever the allocated sample sizes reach this minimum. In practice, it may even be wise to use $n_{\min}$ larger than 2, because of nonresponse or for other practical reasons.

Assuming full response, the data are collected for all units in the selected sample and used to produce estimates (of totals, say) for a set of m survey variables. Let $y_1, y_2, \ldots, y_m$ denote the survey variables. The variance of variable y_j in stratum h is defined as:

$$S_{hj}^2 = \frac{1}{N_h - 1} \sum_{i \in U_h} \left(y_{ij} - \bar{Y}_{hj} \right)^2 \tag{2.4}$$

where y_{ij} is the value of y_j for the i^{th} population unit, and $\bar{Y}_{hj}$ is the population mean for y_j in stratum $h,$ given by

$$\bar{Y}_{hj} = \frac{1}{N_h} \sum_{i \in U_h} y_{ij} = Y_{hj} \big/ N_h \tag{2.5}$$

for $h = 1, \ldots, H$ and $j = 1, \ldots, m.$ The population total Y_j for the j^{th} survey variable is $Y_j = \sum_{h=1}^{H} \sum_{i \in U_h} y_{ij} = \sum_{h=1}^{H} Y_{hj}.$

Under stratified simple random sampling (STSRS), the variance of the Horvitz-Thompson (HT) estimator t_j of the total for the j^{th} survey variable (Cochran 1977) is given by:

$$V\left(t_j\right) = \sum_{h=1}^{H} N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) S_{hj}^2 \tag{2.6}$$

where $t_j = \sum_{h=1}^{H} N_h \big/ n_h \sum_{i \in s_h} y_{ij} = \sum_{h=1}^{H} N_h \bar{y}_{hj},$ $s_h \subset U_h$ is the set of labels of the n_h units sampled in stratum $h,$ and $\bar{y}_{hj}$ is the sample mean in stratum $h.$

Because the values of N_h and S_{hj}^2 are fixed after the strata have been defined, the variance of the HT estimator t_j of the total for the j^{th} survey variable in (2.6) depends only on the sample sizes n_h allocated to the strata. This allocation is important, because it is what enables the survey designer to control the precision of the survey estimates.

In general, when performing the allocation, the survey planner seeks a balance between achieving the desired precision for each of the survey variables of interest and the cost of the survey. The importance and computational complexity of this problem have motivated many contributions, which consider one of the two goals of the allocation problem, as described in Section 1. See for example Kokan (1963), Folks and Antle (1965), Kokan and Khan (1967), Huddleston, Claypool and Hocking (1970), Kish (1976), Bethel (1985, 1989), Chromy (1987), Valliant and Gentle (1997), Khan and Ahsan (2003), García and

Cortez (2006), Kozak (2006), Day (2010), Khan, Ali and Ahmad (2011), Ismail, Nasser and Ahmad (2011), Khan, Ali, Raghav and Bari (2012).

All of the above apply methods based on linear programming theory, convex programming, dynamic programming, multi-objective programming and heuristics to try and solve the multivariate optimal allocation problem. Here we propose two integer programming formulations to tackle the problem.

Formulation A

$$\text{Minimize } \sum_{h=1}^{H} c_h n_h \tag{2.7}$$

$$\text{s.t. } n_{\min} \le n_h \le N_h,\ h = 1, \ldots, H \tag{2.8}$$

$$\sqrt{V\left(t_j\right)} \big/ Y_j \le \text{CV}_j \quad j = 1, \ldots, m \tag{2.9}$$

$$n_h \in Z_+ \quad h = 1, \ldots, H \tag{2.10}$$

where c_h represents the unit level survey cost for sampling from stratum h.

In this formulation, the objective function to be minimized (2.7) corresponds to the overall variable cost budget for the survey (which we denote by C). If the unit level survey costs for sampling from the various strata are unknown or are assumed to be the same, then c_h may all be set to one and the alternative objective function to minimize is $n = \sum_{h=1}^{H} n_h$, namely the overall sample size.

Constraint (2.8) ensures that at least $n_{\min}$ units are allocated to each stratum, and that the sample size will not exceed the population size for the stratum.

Constraint (2.9) ensures that the CV of the HT estimator of total for each survey variable is below a pre-specified threshold CV_j $(j = 1, \ldots, m)$ called target CV. Finally, constraint (2.10) ensures that all the allocated sample sizes are integers.

Note that the constraints (2.9) may be rewritten as:

$$\frac{V(t_j)}{Y_j^{\,2}\ \text{CV}_j^2} \le 1\ ,\ j = 1, \ldots, m. \tag{2.11}$$

Now replacing the numerator in (2.11) by equation (2.6), leads to:

$$\sum_{h=1}^{H} \left(\frac{N_h^2 S_{hj}^2}{n_h Y_j^2 \text{CV}_j^2} - \frac{N_h S_{hj}^2}{Y_j^2 \text{CV}_j^2} \right) \le 1, \quad j = 1, \ldots, m. \tag{2.12}$$

Defining

$$p_{hj} = \frac{N_h\ S_{hj}^2}{Y_j^2\ \text{CV}_j^2} \tag{2.13}$$

the constraints (2.12) may be written as:

$$\sum_{h=1}^{H} \left(\frac{N_h \; p_{hj}}{n_h} - p_{hj} \right) \leq 1, \quad j = 1, \ldots, m. \tag{2.14}$$

Formulation B

$$\text{Minimize} \sum_{j=1}^{m} w_j \frac{1}{Y_j^2} \left[\sum_{h=1}^{H} N_h^2 \; \left(\frac{1}{n_h} - \frac{1}{N_h} \right) \; S_{hj}^2 \right] \tag{2.15}$$

$$\text{s.t. } n_{\min} \leq n_h \leq N_h, \quad h = 1, \ldots, H \tag{2.16}$$

$$\sum_{h=1}^{H} c_h n_h \leq C \tag{2.17}$$

$$n_h \in Z_+ \quad h = 1, \ldots, H \tag{2.18}$$

$$0 < w_j < 1 \;\; \forall j \;\; \text{and} \;\; \sum_{j=1}^{m} w_j = 1 \tag{2.19}$$

where w_j are variable-specific weights, set a priori to represent the relative importance of the survey variables. The variable-specific weights w_j are set by subject matter experts or the survey designers. If they are not specified, equal relative weights could be assigned to all the survey variables considered.

In this formulation, the objective function (2.15) to be minimized corresponds to a weighted sum of the relative variances of the estimates of total for the m survey variables. We use relative variances because different survey variables may be measured in different units, and thus summing variances is not meaningful. Examining (2.15) it is clear that its minimum is achieved when

$$\sum_{j=1}^{m} w_j \frac{1}{Y_j^2} \left[\sum_{h=1}^{H} N_h^2 \; \left(\frac{1}{n_h} \right) \; S_{hj}^2 \right]$$

is minimum, since the last term

$$\sum_{j=1}^{m} w_j \frac{1}{Y_j^2} \left[\sum_{h=1}^{H} N_h^2 \; \left(-\frac{1}{N_h} \right) \; S_{hj}^2 \right]$$

does not depend on the stratum sample sizes. Hence the objective function (2.15) may be rewritten:

$$\text{Minimize} \sum_{j=1}^{m} w_j \frac{1}{Y_j^2} \left(\sum_{h=1}^{H} \frac{N_h^2}{n_h} \; S_{hj}^2 \right). \tag{2.20}$$

Constraint (2.16) is the same as constraint (2.8) applied in Formulation A. Constraint (2.17) ensures that the total variable cost of the survey will not exceed the allocated budget C. Like constraint (2.10) in Formulation A, constraint (2.18) ensures that all the allocated sample sizes are integers. Constraint (2.19) ensures that the importance weights are adequate for aggregating the relative variances of the estimated totals for each of the survey variables.

When the unit level survey costs c_h per stratum are not known or may be assumed to be equal, constraint (2.17) may be replaced by $\sum_{h=1}^{H} n_h \leq n$ where n is the (maximum) overall sample size.

Both formulations A and B present non-linearity: constraint (2.9) or (2.14) in Formulation A, and the objective function in Formulation B. Therefore a first alternative one could use to resolve the non-linearity problem in these two Formulations would be one of the methods of non-linear programming or convex programming (Bazaraa, Sheralli and Shetty 2006; Luenberger and Ye 2008) that can deal with constraints, as for example penalty based methods or multiplier methods, amongst others. Nevertheless, application of such methods tends to produce solutions (sets of samples sizes to allocate in the strata) that, in general, are non-integers. In addition, when such solutions are rounded to obtain feasible sample sizes, there's no guarantee to obtain a global optimum (Wolsey 1998) in terms of minimizing the corresponding objective functions.

Alternatively, given that the solutions (sample sizes) must be integers, one could consider applying integer programming methods, such as *Branch and Bound* (Land and Doig 1960; Wolsey 1998; Wolsey and Nemhauser 1999). However, the non-linearity present in both formulations prevents the immediate application of such methods.

With these issues in mind, in the next section we propose two new formulations for integer programming that circumvent these problems and are equivalent to the Formulation A, defined jointly by (2.7), (2.8), (2.9) and (2.10), and Formulation B, defined jointly by (2.20), (2.16), (2.17), (2.18) and (2.19). More specifically, from the resolution of these new formulations it is possible to obtain integer sample sizes (n_h) for the sample allocation which satisfy the constraints established for each problem and also lead to a global optimum (Wolsey 1998) either for the objective function defined in (2.7), or for the objective function defined in (2.20), respectively.

3 Proposed formulations

From an optimization point of view, solving the problems defined by (2.7) – (2.10) or by (2.16) – (2.20) consists of determining $n_1, n_2, \ldots n_H$ chosen from the sets defined by $A_h = \{n_{\min}, \ldots, N_h\}$, $h = 1, \ldots, H$, that the constraints in each of these problems are satisfied and the corresponding objective function is minimized. As already indicated, a standard minimum sample size per stratum of $n_{\min} = 2$ is considered here to define the sets A_h, but this value may be changed as needed to accommodate specific survey requirements.

Taking this approach, a new formulation may be considered where the decision variables are indicator variables of which elements of the sets A_h $(h = 1, \ldots, H)$ will be chosen. For this purpose, we introduce the binary variable x_{hk} taking the value 1 if the sample size $k \in A_h$ is allocated to stratum h, and value 0 if this sample size is not allocated to stratum $h, h = 1, \ldots, H$.

Considering the formulations previously presented and these new binary variables, we may write two integer programming formulations where the decision variables (i.e., the unknowns to be determined) are of the $0-1$ type, therefore configuring a binary integer programming problem (Wolsey and Nemhauser 1999). The formulation equivalent to Formulation A is given by:

Formulation C

$$\text{Minimize } \sum\nolimits_{h=1}^{H} c_h \left(\sum\nolimits_{k=n_{\min}}^{N_h} k \; x_{hk} \right) \tag{3.1}$$

$$\text{s.t. } \sum\nolimits_{k=n_{\min}}^{N_h} x_{hk} = 1 \;\; \forall h = 1, \ldots, H \tag{3.2}$$

$$\sum\nolimits_{h=1}^{H} N_h \; p_{hj} \left(\sum_{k=n_{\min}}^{N_h} \frac{x_{hk}}{k} \right) - \sum\nolimits_{h=1}^{H} p_{hj} \le 1, \;\; j = 1, \ldots, m \tag{3.3}$$

$$x_{hk} \in \{0, 1\}, \;\; k = n_{\min}, \ldots, N_h, h = 1, \ldots, H. \tag{3.4}$$

In Formulation C, constraint (3.2) ensures that, for each of the strata, there will be exactly one x_{hk} variable taking the value one. This is equivalent to ensuring the choice of only one value k (the sample size) from each set $A_h \; (h = 1, \ldots, H)$. Constraint (3.3) is equivalent to constraint (2.9) or its equivalent (2.14) in Formulation A. This formulation contemplates potentially varying unit survey costs for the various strata. If this is not necessary, the objective function in (3.1) may be redefined as

$$\text{Minimize } \sum\nolimits_{h=1}^{H} \sum\nolimits_{k=n_{\min}}^{N_h} k \; x_{hk}. \tag{3.5}$$

In order to help with the understanding of the proposed formulation, consider the following example.

Example 1: Suppose that there are three population strata $(H = 3)$ with $N_1 = 3, N_2 = 5$ and $N_3 = 4,$ that the unit survey costs are the same across strata (say $c_h = 1 \; \forall h)$ and only one survey variable $(m = 1)$. Formulation C would then look like:

$$\text{Minimize } 1\, x_{11} + 2\, x_{12} + 3\, x_{13} + 1\, x_{21} + 2\, x_{22} + 3\, x_{23} + 4\, x_{24} + 5\, x_{25} + 1\, x_{31} + 2\, x_{32} + 3\, x_{33} + 4\, x_{34} \tag{3.6}$$

$$\text{s.t. } x_{11} + x_{12} + x_{13} = 1 \tag{3.7}$$

$$x_{21} + x_{22} + x_{23} + x_{24} + x_{25} = 1 \tag{3.8}$$

$$x_{31} + x_{32} + x_{33} + x_{34} = 1 \tag{3.9}$$

$$\begin{gathered} N_1 \; p_{11} \left(1\, x_{11} + \frac{1}{2} \; x_{12} + \frac{1}{3} \; x_{13} \right) - p_{11} + N_2 \; p_{21} \left(1\, x_{21} + \frac{1}{2} \; x_{22} + \frac{1}{3} \; x_{23} + \frac{1}{4} \; x_{24} + \frac{1}{5} \; x_{25} \right) - p_{21} + \\ N_3 \; p_{31} \left(1\, x_{31} + \frac{1}{2} \; x_{32} + \frac{1}{3} \; x_{33} + \frac{1}{4} \; x_{34} \right) - p_{31} \le 1 \end{gathered} \tag{3.10}$$

$$x_{11}, x_{12}, x_{13}, x_{21}, x_{22}, x_{23}, x_{24}, x_{25}, x_{31}, x_{32}, x_{33}, x_{34} \in \{0, 1\}. \tag{3.11}$$

Formulation B may also be translated to this new approach of using the binary variables as follows.

Formulation D

$$\text{Minimize } \sum\nolimits_{j=1}^{m} w_j \frac{1}{Y_j^2} \sum\nolimits_{h=1}^{H} \left(\sum\nolimits_{k=n_{\min}}^{N_h} \frac{x_{hk}}{k} \right) N_h^2 S_{hj}^2 \tag{3.12}$$

$$\text{s.t. } \sum\nolimits_{k=n_{\min}}^{N_h} x_{hk} = 1 \;\; \forall h = 1, \ldots, H \tag{3.13}$$

$$\sum\nolimits_{h=1}^{H} c_h \left(\sum\nolimits_{k=n_{\min}}^{N_h} k \; x_{hk} \right) \leq C \tag{3.14}$$

$$x_{hk} \in \{0,1\}, \;\; k = n_{\min}, \ldots, N_h, h = 1, \ldots, H. \tag{3.15}$$

In Formulation D the objective function (3.12) is equivalent to the objective function (2.20). Constraint (3.13) is equivalent to constraint (2.16). Constraint (3.14) is equivalent to constraint (2.17) and ensures that the total variable cost of the survey will not exceed the allocated budget C. In case we do not have information on unit survey costs per strata, or wish to consider that they are the same across the strata, we replace constraint (3.14) by

$$\sum\nolimits_{h=1}^{H} \sum\nolimits_{k=n_{\min}}^{N_h} k \; x_{hk} \leq n. \tag{3.16}$$

In order to illustrate the proposed formulation, consider the following example.

Example 2: Suppose that there are two population strata $(H = 2)$ with $N_1 = 3$ and $N_2 = 4,$ with two survey variables $(m = 2)$, equal unit survey costs for both strata, importance weights w_j equal to ½ for both survey variables and a total sample size of $n = 5.$ Formulation D would then look like:

$$\begin{aligned} \text{Minimize } & \left[x_{11} \frac{N_1^2}{1} + x_{12} \frac{N_1^2}{2} + x_{13} \frac{N_1^2}{3} \right] \frac{1}{2} \left(\frac{S_{11}^2}{Y_1^2} + \frac{S_{12}^2}{Y_2^2} \right) + \\ & \left[x_{21} \frac{N_2^2}{1} + x_{22} \frac{N_2^2}{2} + x_{23} \frac{N_2^2}{3} + x_{24} \frac{N_2^2}{4} \right] \frac{1}{2} \left(\frac{S_{21}^2}{Y_1^2} + \frac{S_{22}^2}{Y_2^2} \right) \end{aligned} \tag{3.17}$$

$$\text{s.t. } x_{11} + x_{12} + x_{13} = 1 \tag{3.18}$$

$$x_{21} + x_{22} + x_{23} + x_{24} = 1 \tag{3.19}$$

$$1\, x_{11} + 2\, x_{12} + 3\, x_{13} + 1\, x_{21} + 2\, x_{22} + 3\, x_{23} + 4\, x_{24} \leq 5 \tag{3.20}$$

$$x_{11}, x_{12}, x_{13}, x_{21}, x_{22}, x_{23}, x_{24} \in \{0,1\}. \tag{3.21}$$

In this paper, these two formulations were resolved applying a method of implicit enumeration called *Branch and Bound*. *Branch and Bound* (Wolsey 1998, Wolsey and Nemhauser 1999) methods obtain the optimal solution for binary integer programming problems efficiently, by considering the resolution of a subset of problems associated with the feasible region for the problem. These methods were developed from the pioneer work of Land and Doig (1960).

The solutions for both Formulations C and D of the approach, labelled BSSM (the initials for Brito, Silva, Semaan and Maculan), were obtained using the R package $Rglpk$. The R code we developed is available on request. The package $Rglpk$ contains a set of procedures that can be applied for solving linear and integer programming problems.

For comparison purposes in the case of our Formulation C, we also considered in our numerical illustrations an algorithm proposed by (Bethel 1985 and 1989) which is available in the R package $SamplingStrata$. This algorithm relies on the Kuhn-Tucker Theorem, and uses the Lagrange multipliers (Bazaraa, et al. 2006). In the case of our Formulation D, we compared our approach with a 'textbook method' proposed in Cochran (1977, Section 5.A.4), as suggested by the Associate Editor.

4 Numerical results

This section provides results for the application of the selected multivariate optimum allocation approaches to a set of population datasets. The approaches considered include:

- The BSSM algorithms developed to solve Formulations C and D provided in Section 3;
- An improved version of Bethel's algorithm (Bethel 1989) developed by Ballin and Barcarolli (2008);
- The textbook method proposed in Cochran (1977, Section 5.A.4).

Eleven population datasets were used for the numerical illustration, but for space considerations, here we report only the results for three of these populations. The three selected populations are described in tables A1 through A6 in Appendix A. Table A1 provides a brief description of each survey population and provides the list of the corresponding survey variables. Table A2 provides information about how each population was stratified prior to determining the optimum allocation. In particular, for the survey dataset called $MunicSw$ the strata had been previously defined. The other two populations were stratified using a stratification algorithm available in the R package $stratification$ or a classic $k-\text{means}$ clustering method available in the $base$ R package.

Table A3 presents the number of population strata (H), the number of survey variables (m), and the population size (N) for each of the populations considered. Tables A4 through A6 provide the population counts, means, and standard deviations per stratum for the survey variables considered in each of the three survey populations considered.

The results of all the numerical experiments reported here were obtained using the R packages and functions mentioned, and using a Windows 7 desktop computer with 24GB of RAM and with eight i7 processors of 3.40GHz. Processing time ranged from miliseconds (for the relatively small $MunicSw$ population) to less than 4 seconds (for the larger $SchoolsNortheast$ population, under formulation C). This demonstrates that the proposed formulations provide a feasible and efficient alternative for multivariate optimum allocation problems of small and medium size, for populations of sizes (N) in thousands and even tens of thousands.

Tables 4.1 through 4.3 provide the target coefficients of variation $\left(\text{CV}_j\right)$ for each of the survey variables, the sample sizes obtained using the algorithm to solve proposed Formulation C $\left(n_{\text{BSSM}}\right)$ and Bethel's algorithm $\left(n_{\text{Bethel}}\right)$, and the achieved coefficients of variation for the estimators of totals of the survey variables considered in each population under the two algorithms compared.

Table 4.1
Results for the *CoffeeFarms* population

	Algorithm for Formulation C				Bethel's Algorithm			
CV_j	n_{BSSM}	$\text{CV}(t_1)$	$\text{CV}(t_2)$	$\text{CV}(t_3)$	n_{Bethel}	$\text{CV}(t_1)$	$\text{CV}(t_2)$	$\text{CV}(t_3)$
(%)		(%)	(%)	(%)		(%)	(%)	(%)
5	**2,545**	1.24	5.00	2.92	2,546	1.23	5.00	2.91
10	**754**	3.30	10.00	7.01	755	3.30	9.99	7.07
15	**347**	5.21	15.00	11.01	349	5.11	14.95	10.85

Table 4.2
Results for the *SchoolsNortheast* population

	Algorithm for Formulation C			Bethel's Algorithm		
CV_j	n_{BSSM}	$\text{CV}(t_1)$	$\text{CV}(t_2)$	n_{Bethel}	$\text{CV}(t_1)$	$\text{CV}(t_2)$
(%)		(%)	(%)		(%)	(%)
2	**1,624**	2.00	1.79	1,628	2.00	1.78
5	**294**	5.00	4.31	299	4.96	4.23
10	**80**	9.93	8.24	83	9.72	8.13

Table 4.3
Results for the *MunicSw* population

	Algorithm for Formulation C					Bethel's Algorithm				
CV_j	n_{BSSM}	$\text{CV}(t_1)$	$\text{CV}(t_2)$	$\text{CV}(t_3)$	$\text{CV}(t_4)$	n_{Bethel}	$\text{CV}(t_1)$	$\text{CV}(t_2)$	$\text{CV}(t_3)$	$\text{CV}(t_4)$
(%)		(%)	(%)	(%)	(%)		(%)	(%)	(%)	(%)
5	**1,527**	2.01	3.88	5.00	4.41	1,529	2.00	3.88	4.99	4.40
10	**761**	3.61	7.27	9.99	8.77	763	3.60	7.25	9.97	8.75
15	**439**	5.01	10.22	14.98	13.07	441	4.95	10.16	14.94	13.03

As expected, in all cases the sample sizes obtained by solving Formulation C were smaller than (bold) or equal to those obtained using Bethel's algorithm. However, the improvements were generally not substantial. Nevertheless the proposed algorithm managed to improve upon the current best method in the nine scenarios considered (three populations times three levels for the target CVs). The improvements appeared to be a bit larger for the *SchoolsNortheast* Population, where the number of strata is also larger. Similar results (not shown here for conciseness but available from the authors on request) were obtained for the other eight populations considered in an initial version of the paper.

Tables 4.4 to 4.6 provide the results of applying Formulation D and the textbook method proposed in Cochran (1977, Section 5.A.4) to the same three survey populations. Now the goal is to minimize the weighted relative variance of the HT estimates of total, while keeping the overall sample size or cost. The first line in each of these tables contains the total sample sizes considered for the allocation. These sample sizes correspond to sampling fractions of 10%, 20% and 30% of the corresponding population sizes (N)

respectively, as indicated in the second line in each of the tables. The subsequent lines provide the allocation of the total sample into the strata, the coefficients of variation achieved for the HT estimates of totals of the survey variables considering the allocation, and the sum of the coefficients of variation $(\Sigma \mathrm{CV}(t_i))$, which is a summary measure of efficiency across all survey variables.

The importance weights were taken as equal across all survey variables, and the unit survey costs were taken as equal across all strata, in each population, for these applications.

Table 4.4
Results for the *CoffeeFarms* population

n	**2.047**		**4.094**		**6.142**	
Sampling fraction	**10%**		**20%**		**30%**	
Result	**BSSM-D**	**Textbook**	**BSSM-D**	**Textbook**	**BSSM-D**	**Textbook**
n_1	1,174	1,124	2,483	2,340	3,792	3,625
n_2	662	737	1,400	1,544	2,139	2,306
n_3	211	186	211	210	211	211
$\mathrm{CV}(t_1)$	**1.02**	1.14	**0.62**	0.62	**0.42**	0.42
$\mathrm{CV}(t_2)$	**5.78**	5.79	**3.62**	3.65	**2.61**	2.63
$\mathrm{CV}(t_3)$	**2.86**	2.98	**1.73**	1.73	1.19	**1.17**
$\Sigma\mathrm{CV}(t_i)$	**9.66**	9.91	**5.97**	6.00	**4.22**	4.22

Table 4.5
Results for the *SchoolsNortheast* population

n	**7,508**		**15,017**		**22,525**	
Sampling fraction	**10%**		**20%**		**30%**	
Result	**BSSM-D**	**Textbook**	**BSSM-D**	**Textbook**	**BSSM-D**	**Textbook**
n_1	82	58	82	60	82	66
n_2	36	33	62	53	53	62
n_3	7	6	7	6	7	6
n_4	206	214	465	433	771	611
n_5	1,083	1,000	2,091	1,962	2,671	2,121
n_6	447	452	891	914	1,428	1,436
n_7	361	371	711	750	1,182	1,175
n_8	2,995	2,989	5,963	6,055	9,088	9,634
n_9	976	1,023	1,965	2,069	3,078	3,229
n_{10}	399	419	800	849	1,331	1,338
n_{11}	797	813	1,742	1,647	2,596	2,612
n_{12}	119	130	238	219	238	235
$\mathrm{CV}(t_1)$	**0.86**	0.98	**0.54**	0.69	**0.39**	0.54
$\mathrm{CV}(t_2)$	0.73	**0.72**	**0.47**	0.47	0.35	**0.34**
$\Sigma\mathrm{CV}(t_i)$	**1.59**	1.70	**1.01**	1.16	**0.74**	0.88

Table 4.6
Results of formulation D for the *MunicSw* population

n *Sampling fraction*	290 10%		579 20%		869 30%	
Result	**BSSM-D**	**Textbook**	**BSSM-D**	**Textbook**	**BSSM-D**	**Textbook**
n_1	67	59	134	118	202	182
n_2	68	77	136	153	206	233
n_3	40	35	80	70	120	107
n_4	58	47	116	93	171	128
n_5	32	43	65	85	97	129
n_6	16	21	31	43	47	65
n_7	9	8	17	17	26	25
$\text{CV}(t_1)$	5.93	**5.40**	4.01	**3.61**	3.10	**2.75**
$\text{CV}(t_2)$	12.53	**12.24**	8.36	**8.12**	6.36	**6.14**
$\text{CV}(t_3)$	**19.49**	20.19	**12.46**	13.01	**8.95**	9.56
$\text{CV}(t_4)$	**16.91**	17.45	**10.85**	11.27	**7.84**	8.30
$\Sigma\text{CV}(t_i)$	**54.86**	55.28	**35.68**	36.01	**26.25**	26.75

As expected, in all three cases the sum of the coefficients of variation obtained by solving Formulation D were smaller than (bold) those obtained using the textbook algorithm. However, the textbook algorithm provided smaller CVs for some of the survey variables, in particular for the *MunicSw* population. The improvements were generally not very large, but again were slightly larger for the *SchoolsNortheast* population. In this comparison, however, the allocations are quite different between the two methods.

5 Final remarks

In this paper we provided two new formulations leading to the achievement of the global minimum in multivariate optimum allocation problems. These exact integer programming formulations can be efficiently implemented using off the shelf free software (namely the *Rglpk* R package). In addition, the proposed formulations enable the definition of minimum sample sizes per strata, something which is clearly of interest in practice to avoid allocations with sample sizes less than 2, for example, which would lead to difficulties regarding variance estimation. Such minimum sample sizes may be set at larger values (say 5, 10, 30 or some other number) to ensure that the samples are large enough to tolerate some nonresponse or to ensure estimation is feasible for each stratum, if the strata are used as estimation domains.

The proposed approach improves upon the existing methods by tackling the allocation problem directly, and dealing with the non-linearity of either the objective function or the constraints, as well as the requirement that the solution provides only integer sample sizes for the strata. In the literature, previously existing methods tackle the problem with approaches which are not guaranteed to reach the global optimum, or that produce real-valued allocations that must be rounded to integer-values.

In practice, finding real-valued allocations is not a big problem, unless the stratum population sizes N_h are very small or when there is a very large number of strata. In the first case, sampling one unit more, or less, can make a big change in the sampling fractions, which can cause some large impacts in the variances. In the second case, rounding the allocated sample sizes can make a difference in the total sample size n. When all the stratum population sizes N_h are relatively large, and the number of strata is reasonable, rounding non-integer sample sizes will not create a problem.

In this paper we carried out some limited numerical work, aimed essentially at demonstrating the feasibility of the proposed approach. The results obtained using Formulation C of the proposed approach are comparable to those achieved using the Bethel method, while providing integer-valued allocations that correspond to the global optimum. But given that only little differences were found between the two methods (BSSM and Bethel) in the applications considered, there may be little incentive to move to the BSSM method. The results obtained under Formulation D showed modest improvements over the textbook method used in the comparison.

Further research is needed to test the approach for larger problems and to assess its merits compared to other methods under other practical scenarios. An important advantage of the proposed approach is that both formulations can be implemented using off the shelf software, as indicated.

Acknowledgements

This research was supported by FAPERJ. Research Grant E-26/111.947/2012.

Appendix A

Description of the survey populations considered in the numerical experiment

Table A1
Description of the populations

Population	**Description**	**Survey Variables (y)**
CoffeeFarms	Coffee farms in the state of Paraná, Brazil, from 1996 Agricultural Census.	Number of Coffee Trees Total Farm Area Coffee Production
SchoolsNortheast	Data from the 2012 census of schools, by school, for schools in the Northeast region of Brazil.	Number of classrooms Number of employees
MunicSw	Information about Swiss municipalities from the package *SamplingStrata*.	Area of Farming Industrial Area Number of Households Population

Table A2
Stratification of the populations

Population	Stratification
CoffeeFarms	Stratified considering the Number of Coffee Trees variable, using the Kozak algorithm available in the *Stratification* package.
SchoolsNortheast	Twelve strata were formed considering: school type (4 classes), and school size - number of students (3 classes). School size stratification was performed using $k -$ means clustering algorithm within each school type.
MunicSw	This population is available from the *SamplingStrata* package and the strata correspond to regions of Switzerland.

Table A3
Number of strata, number of survey variables and total size for the survey populations considered

Population	H	m	N
CoffeeFarms	3	3	20,472
SchoolsNortheast	12	2	75,084
MunicSw	7	4	2,896

Table A4
Population summaries per stratum $- CoffeeFarms$

Summary	Stratum		
	$h = 1$	$h = 2$	$h = 3$
N_h	17,821	2,440	211
$\bar{Y}_{1h}$	4,291	26,688	218,712
$\bar{Y}_{2h}$	22	84	488
$\bar{Y}_{3h}$	2,671	13,204	129,033
S_{h1}	2,873	15,541	193,366
S_{h2}	69	262	583
S_{h3}	4,611	24,704	200,447

Table A5
Population summaries per stratum $-$ *SchoolsNortheast*

Stratum	N_h	$\bar{Y}_{1h}$	$\bar{Y}_{2h}$	S_{h1}	S_{h2}
$h = 1$	82	45.1	54.0	309.2	24.9
$h = 2$	63	23.9	146.3	14.4	92.6
$h = 3$	7	80.9	700.4	29	342.5
$h = 4$	783	16.2	95.7	6.4	49.5
$h = 5$	2,676	10.9	57.7	21.6	23.7
$h = 6$	3,958	6.1	26.7	4.2	17.9
$h = 7$	2,172	13.6	76.8	5.7	27.9
$h = 8$	45,243	2.5	9.3	3	8.8
$h = 9$	9,674	7.7	38.0	3.2	17.9
$h = 10$	1,743	17.3	49.1	9.2	36.7
$h = 11$	8,445	7.3	15.3	4.1	13.5
$h = 12$	238	37.7	140.8	18.4	88.9

Table A6
Population summaries per stratum – *MunicSw*

Summary	Statum						
	$h = 1$	$h = 2$	$h = 3$	$h = 4$	$h = 5$	$h = 6$	$h = 7$
N_h	589	913	321	171	471	186	245
$\bar{Y}_{1h}$	262.5	367.2	262.7	438.0	429.5	668.9	47.0
$\bar{Y}_{2h}$	5.5	5.3	9.7	13.3	7.9	11.0	4.1
$\bar{Y}_{3h}$	963.9	782.1	1,345.2	3,319.1	906.0	1,465.2	550.7
$\bar{Y}_{4h}$	2,252.5	1,839.4	3,099.5	7,297.7	2,226.0	3,675.8	1,252.4
S_{h1}	220.5	342.4	173.2	290.2	414.2	568.7	65.3
S_{h2}	15.1	13.0	19.4	29.7	14.9	15.5	8.2
S_{h3}	4,600.9	2,794.7	5,003.5	14,610.0	2,178.6	2,802.1	1,197.5
S_{h4}	9,540.3	5,621.6	9,764.5	28,589.4	4,759.4	5,914.5	2,514.9

References

Ballin, M., and Barcaroli, G. (2008). Optimal stratification of sampling frames in a multivariate and multidomain sample design. *Contributi ISTAT*, 10.

Bazaraa, M.S., Sherali, H.D. and Shetty, C.M. (2006). *Nonlinear Programming: Theory and Algorithms*. New York: John Wiley & Sons, Inc, Third Edition.

Bethel, J. (1985). An optimum allocation algorithm for multivariate surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 209-212.

Bethel, J. (1989). Sample allocation in multivariate surveys. *Survey Methodology*, 15, 1, 47-57.

Chromy, J. (1987). Design optimization with multiple objectives. *Proceedings of the Survey Research Methods Section,* American Statistical Association, 194-199.

Cochran, W.G. (1977). *Sampling Techniques*. Third Edition-Wiley.

Day, C.D. (2010). A multi-objective evolutionary algorithm for multivariate optimal allocation. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Folks, J.L., and Antle, C.E. (1965). Optimum allocation of sampling units to strata when there are R responses of interest. *Journal of the American Statistical Association*, 60 (309), 225-233.

García, J.A.D., and Cortez, L.U. (2006). Optimum allocation in multivariate stratified sampling: Multi-objective programming. *Comunicaciones Del Cimat*, no I-06-07/28-03-2006.

Huddleston, H.F., Claypool, P.L. and Hocking, R.R. (1970). Optimal sample allocation to strata using convex programming. *Journal of the Royal Statistical Society, Series C*, 19 (3).

Ismail, M.V., Nasser, K. and Ahmad, Q.S. (2011). Solution of a multivariate stratified sampling problem through Chebyshev's Goal programming. *Pakistan Journal of Statistics and Operation Research*, vol. vii, 1, 101-108.

Khan, M.G.M., and Ahsan, M.J. (2003). A note on optimum allocation in multivariate stratified sampling. *The South Pacific Journal of Natural Science*, 21, 91-95.

Khan, M.F., Ali, I. and Ahmad, Q.S. (2011). Chebyshev approximate solution to allocation problem in multiple objective surveys with random costs. *American Journal of Computational Mathematics*, 1, 247-251.

Khan, M.F., Ali, I., Raghav, Y.S. and Bari, A. (2012). Allocation in multivariate stratified surveys with non-linear random cost function. *American Journal of Operations Research*, 2, 100-105.

Kish, L. (1976). Optima and proxima in linear sample designs. *Journal of the Royal Statistical Society*, *Series A*, 139 (1), 80-95.

Kokan, A.R. (1963). Optimum allocation in multivariate surveys. *Journal of the Royal Statistical Society*, *Series A*, 126 (4), 557-565.

Kokan, A.R., and Khan, S. (1967). Optimum allocation in multivariate surveys: An analytical solution. *Journal of the Royal Statistical Society*, *Series B*, 29 (1), 115-125.

Kozak, M. (2006). Multivariate sample allocation: Application of random search method. *Statistics in Transition*, 7 (4), 889-900.

Land, A.H., and Doig, A.G. (1960). An Automatic method for solving discrete programming problems. *Econometrica*, 28 (3), 497-520.

Lohr, S.L. (2010). *Sampling: Design and Analysis,* Second edition. Brooks/Cole, Cengage Learning.

Luenberger, D.G., and Ye, Y. (2008*). Linear and Non-Linear Programming*, Third Edition. Springer.

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

Valliant, R., and Gentle, J.E. (1997). An application of mathematical programming to sample allocation. *Computational Statistics & Data Analysis*, 25, 337-360.

Wolsey, L.A. (1998). *Integer Programming*. Wiley-Interscience Series in Discrete Mathematics and Optimization.

Wolsey, L.A., and Nemhauser, G.L. (1999). *Integer and Combinatorial Optimization*. Wiley-Interscience Series in Discrete Mathematics and Optimization.

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following people who have provided help or served as referees for one or more papers during 2015.

P. Aronow, *Yale University*
J.-F. Beaumont, *Statistics Canada*
E. Berg, *Iowa State University*
Y. G. Berger, *University of Southampton*
C. Bocci, *Statistics Canada*
J. Breidt, *Colorado State University*
M. Brick, *Westat Inc.*
D. Cantor, *Westat Inc.*
P.J. Cantwell, *U.S. Bureau of the Census*
S. Casanova, *Université de Toulouse*
C. Casas-Cordero, *Pontificia Universidad Catolica de Chile*
G. Chauvet, *ENSAI*
S. Chen, *University of Oklahoma*
Y. Cheng, *U.S. Census Bureau*
R. Clark, *University of Wollongong*
M. Couper, *Survey Research Center, University of Michigan*
J. Dever, *Research Triangle Institute*
M. S. Diallo, *Westat Inc.*
J.L. Eltinge, *U.S. Bureau of Labor Statistics*
W.A. Fuller, *Iowa State University*
J. Gambino, *Statistics Canada*
C. Girard, *Statistics Canada*
T. Goicon, *University of Navarra*
D. Haziza, *Université de Montréal*
B. Hulliger, *University of Applied Sciences Northwestern Switzerland*
P. Joyce, *U.S. Census Bureau*
D. Judkins, *Abt Associates*
M. Karlberg, *Eurostat*
J. Kim, *Iowa State University*
P. Kott, *Research Triangle Institute*
P. Lahiri, *JPSM, University of Maryland*
P. Lavallée, *Statistics Canada*
H. Lee, *Westat Inc.*
P. Lundquist, *Statistics Sweden*
P. Lynn, *University of Essex*
H. Mantel, *Statistics Canada*
A.L. McCutcheon, *University of Nebraska*
C.H. McLaren, *Office for national Statistics*
D.N. da Silva, *Universidade Federal do Rio Grande do Norte*
J. Opsomer, *Colorado State University*
D. Pavlopoulos, *VU University Amsterdam*
D. Pfeffermann, *Hebrew University*
J.N.K. Rao, *Carleton University*
J. Reiter, *Duke University*
L.-P. Rivest, *University of Laval*
A. Ruiz-Gazen, *Toulouse School of Economics*
N. Salvati, *University of Pisa*
F. Scheuren, *National Opinion Research Center*
B. Schouten, *Statistics Netherlands & Utrecht University*
P.L.N.D. Silva, *Escola Nacional de Ciências Estatísticas*
M. Simard, *Statistics Canada*
A. Singh, *Statistics Canada*
P.A. Smith, *University of Southampton*
D. Steel, *University of Wollongong*
P. Steel, *US Census Bureau*
M. Thompson, *University of Waterloo*
S. Thompson, *Simon Fraser University*
Y. Tillé, *Université de Neuchâtel*
M. Torabi, *University of Manitoba*
D. Toth, *Bureau of Labor Statistics*
R. Valliant, *University of Michigan*
J. van den Brakel, *Statistics Netherlands and Maastricht University*
F. Verret, *Statistics Canada*
C. Wu, *University of Waterloo*
G. Yan, *University of Virginia*
W. Yung, *Statistics Canada*
A. Zaslavsky, *Harvard University*
L.-C. Zhang, *Statistics Norway & U. of Southampton*

Acknowledgements are also due to those who assisted during the production of the 2015 issues: Céline Ethier and Nick Budko of International Cooperation and Corporate Statistical Methods Division; Joana Bérubé of Business Survey Methods Division; the team from Dissemination Division, in particular: Chantal Chalifoux, Jacqueline Luffman, Kathy Charbonneau, Joseph Prince et Darquise Pellerin as well as our partners in the Communications Division.

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 31, No. 2, 2015

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 31, No. 3, 2015

All inquires about submissions and subscriptions should be directed to jos@scb.se

The Canadian Journal of Statistics **La revue canadienne de statistique**

CONTENTS TABLE DES MATIÈRES

Volume 43, No. 3, September/septembre 2015

The Canadian Journal of Statistics **La revue canadienne de statistique**

CONTENTS TABLE DES MATIÈRES

Volume 43, No. 4, December/décembre 2015

GUIDELINES FOR MANUSCRIPTS

Authors are invited to submit their articles in English or French in electronic form to the Editor, (statcan.smj-rte.statcan@canada.ca). Before submitting the article, please examine a recent issue of *Survey Methodology* (Vol. 39, No. 1 and onward) as a guide and note particularly the points below. Articles must be submitted in machine-readable form, preferably in Word with MathType for the mathematical expressions. A pdf or paper copy may be required for formulas and figures.

1. Layout

1.1 Documents should be typed entirely double spaced with margins of at least 1½ inches on all sides.
1.2 The documents should be divided into numbered sections with suitable verbal titles.
1.3 The name (fully spelled out) and address of each author should be given as a footnote on the first page of the manuscript.
1.4 Acknowledgements should appear at the end of the text.
1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

3.1 Avoid footnotes, abbreviations, and acronyms.
3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “exp(·)” and “log(·)”, etc.
3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered with arabic numerals on the right if they are to be referred to later. Use a two-level numbering system based on the section of the paper. For example, equation (4.2) is the second important equation in section 4.
3.4 Write fractions in the text using a solidus.
3.5 Distinguish between ambiguous characters, (e.g., w, ω; o, O, 0; l, 1).
3.6 If possible, avoid using bold characters in formulae.

4. Figures and Tables

4.1 All figures and tables should be numbered with arabic numerals, with titles that are as self explanatory as possible, at the bottom for figures and at the top for tables. Use a two-level numbering system based on the section of the paper. For example, table 3.1 is the first table in section 3.
4.2 A detailed textual description of figures may be required for accessibility purposes if the message conveyed by the image is not sufficiently explained in the text.

5. References

5.1 References in the text should be cited with authors’ names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, page 164).
5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

6. Short Notes

6.1 Documents submitted for the short notes section must have a maximum of 3,000 words.