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

Survey Methodology 41-1

Release date: June 29, 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 infostats@statcan.gc.ca

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

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

Depository Services Program

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

Standards of service to the public

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

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

June 2015

•

Volume 41

•

Number 1

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

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

EDITORIAL POLICY

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

Submission of Manuscripts

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

Survey Methodology

A Journal Published by Statistics Canada

Volume 41, Number 1, June 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.
∞

Survey Methodology, June 2015
Vol. 41, No. 1, pp. 1-19
Statistics Canada, Catalogue No. 12-001-X

Small area estimation under a Fay-Herriot model with preliminary testing for the presence of random area effects

Isabel Molina, J.N.K. Rao and Gauri Sankar Datta[1]

Abstract

A popular area level model used for the estimation of small area means is the Fay-Herriot model. This model involves unobservable random effects for the areas apart from the (fixed) linear regression based on area level covariates. Empirical best linear unbiased predictors of small area means are obtained by estimating the area random effects, and they can be expressed as a weighted average of area-specific direct estimators and regression-synthetic estimators. In some cases the observed data do not support the inclusion of the area random effects in the model. Excluding these area effects leads to the regression-synthetic estimator, that is, a zero weight is attached to the direct estimator. A preliminary test estimator of a small area mean obtained after testing for the presence of area random effects is studied. On the other hand, empirical best linear unbiased predictors of small area means that always give non-zero weights to the direct estimators in all areas together with alternative estimators based on the preliminary test are also studied. The preliminary testing procedure is also used to define new mean squared error estimators of the point estimators of small area means. Results of a limited simulation study show that, for small number of areas, the preliminary testing procedure leads to mean squared error estimators with considerably smaller average absolute relative bias than the usual mean squared error estimators, especially when the variance of the area effects is small relative to the sampling variances.

Key Words: Area level model; Empirical best linear unbiased predictor; Mean squared error; Preliminary testing; Small area estimation.

1 Introduction

A basic area-level model, called the Fay-Herriot (FH) model, is often used to obtain efficient estimators of area means when the sample sizes within areas are small. This model involves unobservable area random effects, and the empirical best linear unbiased predictor (EBLUP) of a small area mean is obtained by estimating the associated random effect. The EBLUP is a weighted combination of a direct area-specific estimator and a regression-synthetic estimator that uses all the data. An estimator of the mean squared error (MSE) of the EBLUP was obtained first by Prasad and Rao (1990) using a moment estimator of the random effects variance and later by Datta and Lahiri (2000) for the restricted maximum likelihood (REML) estimator of the variance. Rao (2003, Chapter 7) gives a detailed account of EBLUPs and their MSE estimators for the FH model.

Sometimes the observed data do not support the inclusion of the area effects in the model. Excluding the area effects leads to the regression-synthetic estimator. Using this idea, Datta, Hall and Mandal (2011) proposed to do a preliminary test for the presence of the area random effects at a specified significance level, and then to define the small area estimator depending on the result of the test. If the null hypothesis of no area random effects is not rejected, the model without the area effects is considered to estimate the small area means, i.e., the regression-synthetic estimator is used. If the null hypothesis is rejected, the usual EBLUP under the FH model with area effects is used. Datta et al. (2011) remarked that the above preliminary test estimator (PTE) could lead to significant efficiency gains over the EBLUP, particularly

1. Isabel Molina, Department of Statistics, University Carlos III de Madrid, C/Madrid 126, 28903 Getafe (Madrid), Spain and Instituto de Ciencias Matemáticas (ICMAT), Madrid, Spain. E-mail: isabel.molina@uc3m.es; J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, Canada; Gauri Sankar Datta, Department of Statistics, University of Georgia, Athens, U.S.A.

when the number of small areas is only modest in size. For preliminary testing, they considered a normality-based test as well as a bootstrap test that avoids the normality assumption.

When the estimated area effects variance is zero, the EBLUP becomes automatically the regression-synthetic estimator. However, the estimated MSE obtained by Prasad and Rao (1990) or Datta and Lahiri (2000) does not reduce to the estimated MSE of the regression-synthetic estimator. Thus, the usual MSE estimators are biased for small random effects variance. For this reason, we propose MSE estimators of the EBLUP based on the preliminary testing procedure. If the random effects variance is not significant according to the test, we consider the MSE estimator of the synthetic estimator. Otherwise, we consider the usual MSE estimators of the EBLUP.

The EBLUP attaches zero weight to the direct estimates for all areas when the estimated area effects variance is zero. On the other hand, survey practitioners often prefer to attach a strictly positive weight to the direct estimates, because the latter make use of the available area-specific unit level data and also incorporate the sampling design. Li and Lahiri (2010) introduced an adjusted maximum likelihood (AML) estimator of the variance of random effects that is always positive and therefore leads to EBLUPs giving strictly positive weights to direct estimators. As we shall see, a price is paid in terms of bias when using the EBLUP based on the AML estimator. We propose here alternative small area estimators that always give a positive weight to the direct estimators but with a smaller bias.

This paper studies empirically the properties of PTEs of small area means, in comparison with the usual EBLUPs and other proposed estimators. In particular, we study the choice of the significance level for the area estimates and for the MSE estimates based on the preliminary test (PT). EBLUPs based on the AML estimator of the random effects variance of Li and Lahiri (2010), which give non-zero weights to the direct estimators in all areas, are also studied and compared to PT versions of AML (PT-AML). Different MSE estimators of these PT-AML estimators are also studied with respect to relative bias. Based on simulation results, the EBLUPs and the associated MSE estimators that performed well are recommended. Finally, coverage and length of normality-based prediction intervals, obtained using the EBLUPs and the associated MSE estimators, are examined.

The paper is organized as follows. Section 2 describes the FH model and the EBLUPs of small area means. Section 3 comments on MSE estimation. PTEs of small area means and MSE estimators based on the PT are introduced in Section 4. Section 5 describes small area estimators and associated MSE estimators under AML estimation of the area effects variance. Alternative estimators that also attach positive weights to direct estimators together with proposed MSE estimators are introduced in Section 6. Section 7 reports the results of the simulation study. Finally, Section 8 gives some concluding remarks.

2 Estimation of small area means

Consider a population partitioned into m areas and let θ_i be the mean of the variable of interest for area $i, i = 1, \ldots, m$. We assume that a sample is drawn independently from each area. Let y_i be a design-unbiased direct estimator of θ_i obtained using survey data from the sampled area i. Direct estimators are very inefficient for areas with small sample sizes. We study small area estimation under an area level model, in which the values of area level covariates are available for all areas. The basic model of this type is the Fay-Herriot model, introduced by Fay and Herriot (1979), to estimate per capita income for small

places in the United States. This model consists of two parts. The first part assumes that direct estimators, y_i, of small area means, θ_i, are design unbiased, satisfying

$$y_i = \theta_i + e_i, \quad e_i \overset{\text{ind}}{\sim} N(0, D_i), \quad i = 1, \ldots, m. \tag{2.1}$$

Here, the sampling variance $D_i = \text{Var}(y_i|\theta_i)$ is assumed to be known for all areas $i = 1, \ldots, m$. In practice, the D_i's are ascertained from external sources or by smoothing the estimated sampling variances using a generalized variance function method (Fay and Herriot 1979).

In the second part, the Fay-Herriot model treats θ_i as random and assumes that a p-vector of area level covariates, $\mathbf{x}_i$, linearly related to θ_i, is available for each area i, i.e.,

$$\theta_i = \mathbf{x}_i'\boldsymbol{\beta} + v_i, \quad v_i \overset{\text{iid}}{\sim} N(0, A), \quad i = 1, \ldots, m, \tag{2.2}$$

where v_i is the random effect of area i, assumed to be independent of e_i and $A \geq 0$ is the variance of the random effects. Observe that marginally,

$$y_i \overset{\text{ind}}{\sim} N(\mathbf{x}_i'\boldsymbol{\beta}, D_i + A), \quad i = 1, \ldots, m. \tag{2.3}$$

Letting $\mathbf{y} = (y_1, \ldots, y_m)'$, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_m)'$ and $\mathbf{D} = \text{diag}(D_1, \ldots, D_m)$, model (2.3) may be expressed in matrix notation as $\mathbf{y} \sim N\{\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}(A)\}$ with $\boldsymbol{\Sigma}(A) = \mathbf{D} + A\mathbf{I}_m$, where $\mathbf{I}_m$ denotes the $m \times m$ identity matrix. If A is known, the componentwise best linear unbiased predictor (BLUP) of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)'$ is given by

$$\tilde{\boldsymbol{\theta}}(A) = \left(\tilde{\theta}_1(A), \ldots, \tilde{\theta}_m(A)\right)' = \mathbf{X}\tilde{\boldsymbol{\beta}}(A) + A\boldsymbol{\Sigma}^{-1}(A)\{\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}(A)\}, \tag{2.4}$$

where

$$\begin{aligned} \tilde{\boldsymbol{\beta}}(A) &= \{\mathbf{X}'\boldsymbol{\Sigma}^{-1}(A)\mathbf{X}\}^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}(A)\mathbf{y} \\ &= \left\{\sum_{i=1}^{m}(A + D_i)^{-1}\mathbf{x}_i\mathbf{x}_i'\right\}^{-1}\sum_{i=1}^{m}(A + D_i)^{-1}\mathbf{x}_i y_i \end{aligned} \tag{2.5}$$

is the weighted least squares (WLS) estimator of $\boldsymbol{\beta}$. In practice, however, A is not known. Substituting a consistent estimator $\hat{A}$ for A in the BLUP (2.4), we get the EBLUP given by

$$\hat{\boldsymbol{\theta}} = \left(\hat{\theta}_1, \ldots, \hat{\theta}_m\right)' = \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{A}\hat{\boldsymbol{\Sigma}}^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right), \tag{2.6}$$

where $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\hat{A})$ and $\hat{\boldsymbol{\Sigma}} = \mathbf{D} + \hat{A}\mathbf{I}_m$. For the i^{th} area, the EBLUP of θ_i can be expressed as a convex linear combination of the regression-synthetic estimator $\mathbf{x}_i'\hat{\boldsymbol{\beta}}$ and the direct estimator y_i, as

$$\hat{\theta}_i = B_i(\hat{A})\mathbf{x}_i'\hat{\boldsymbol{\beta}} + \{1 - B_i(\hat{A})\}y_i, \tag{2.7}$$

where the weight attached to the regression-synthetic estimator $\mathbf{x}_i'\hat{\boldsymbol{\beta}}$ is given by $B_i(\hat{A})$, where $B_i(A) = D_i/(A + D_i)$. Observe that the weight increases with the sampling variance D_i. Thus, when the direct estimator is not reliable, i.e., D_i is large as compared with the total variance $\hat{A} + D_i$, more weight is attached to the regression-synthetic estimator $\mathbf{x}_i'\hat{\boldsymbol{\beta}}$. On the other hand, when the direct estimator is efficient, D_i is small relative to $\hat{A} + D_i$, and then more weight is given to the direct estimator y_i.

Several estimators of A have been proposed in the literature including moment estimators without normality assumption, ML estimator and restricted (or residual) ML estimator (REML) estimator. The ML estimator of A is $\hat{A}_{\mathrm{ML}} = \max(0, \hat{A}^*_{\mathrm{ML}})$, where $\hat{A}^*_{\mathrm{ML}}$ can be obtained by maximizing the profile likelihood function given by

$$L_P(A) = c\,|\boldsymbol{\Sigma}(A)|^{-1/2} \exp\left\{-\frac{1}{2}\mathbf{y}'\mathbf{P}(A)\,\mathbf{y}\right\},$$

where c denotes a generic constant and

$$\mathbf{P}(A) = \boldsymbol{\Sigma}^{-1}(A) - \boldsymbol{\Sigma}^{-1}(A)\,\mathbf{X}\{\mathbf{X}'\boldsymbol{\Sigma}^{-1}(A)\,\mathbf{X}\}^{-1}\,\mathbf{X}'\boldsymbol{\Sigma}^{-1}(A).$$

The REML estimator of A is $\hat{A}_{\mathrm{RE}} = \max(0, \hat{A}^*_{\mathrm{RE}})$, where $\hat{A}^*_{\mathrm{RE}}$ is obtained by maximizing the restricted/residual likelihood, given by

$$L_{\mathrm{RE}}(A) = c\,|\mathbf{X}'\boldsymbol{\Sigma}^{-1}(A)\,\mathbf{X}|^{-1/2}\,|\boldsymbol{\Sigma}(A)|^{-1/2} \exp\left\{-\frac{1}{2}\mathbf{y}'\mathbf{P}(A)\,\mathbf{y}\right\}.$$

In this paper, we focus on the REML estimator $\hat{A}_{\mathrm{RE}}$ which is frequently used in practice, and we denote by $\hat{\boldsymbol{\theta}}_{\mathrm{RE}} = (\hat{\theta}_{\mathrm{RE},1}, \ldots, \hat{\theta}_{\mathrm{RE},m})'$ the EBLUP given in (2.6) obtained with $\hat{A} = \hat{A}_{\mathrm{RE}}$.

3 Mean squared error

Note that the BLUP $\tilde{\theta}_i(A)$ of the small area mean θ_i is a linear function of $\mathbf{y}$. Hence, its MSE can be easily calculated and it is given by the sum of two terms:

$$\mathrm{MSE}\{\tilde{\theta}_i(A)\} = g_{1i}(A) + g_{2i}(A),$$

where $g_{1i}(A)$ is due to the estimation of the random area effect v_i and $g_{2i}(A)$ is due to the estimation of the regression parameter $\boldsymbol{\beta}$, with

$$\begin{aligned} g_{1i}(A) &= D_i\{1 - B_i(A)\}, \\ g_{2i}(A) &= B_i^2(A)\,\mathbf{x}_i'\{\mathbf{X}'\boldsymbol{\Sigma}^{-1}(A)\,\mathbf{X}\}^{-1}\,\mathbf{x}_i. \end{aligned}$$

However, the EBLUP $\hat{\theta}_i$ given in (2.7) is not linear in $\mathbf{y}$ due to the estimation of the random effects variance A. Using a moments estimator of A, Prasad and Rao (1990) obtained a second order correct approximation for the MSE of the EBLUP. Later, Datta and Lahiri (2000) and Das, Jiang and Rao (2004)

obtained second order correct MSE approximations under ML and REML estimation of A. When using the REML estimator of A, their approximation to the MSE, for large m, is given by

$$\text{MSE}\left(\hat{\theta}_{\text{RE},i}\right) = g_{1i}(A) + g_{2i}(A) + g_{3i}(A) + o(m^{-1}), \tag{3.1}$$

where

$$g_{3i}(A) = B_i^2(A)\frac{V_{\text{RE}}(A)}{A + D_i} \text{ and } V_{\text{RE}}(A) = \frac{2}{\sum_{i=1}^{m}(A + D_i)^{-2}}.$$

Note that as $m \to \infty, g_{1i}(A) = O(1), g_{2i}(A) = O(m^{-1})$ and $g_{3i}(A) = O(m^{-1})$, so $g_{1i}(A)$ is the leading term in the MSE for large m. However, for small $A, g_{1i}(A)$ is approximately zero and then $g_{3i}(A)$ might be the leading term for small m. For example, taking only one covariate $(p = 1)$ with constant values $x_i = 1$ and constant sampling variances $D_i = D, i = 1, \ldots, m$ and letting $A = 0$, we obtain $g_{1i}(0) = 0, g_{2i}(0) = D/m$ and $g_{3i}(0) = 2D/m$; that is, $g_{3i}(0)$ is twice as large as $g_{2i}(0)$.

Datta and Lahiri (2000) obtained an estimator of the MSE of the EBLUP $\hat{\theta}_{\text{RE},i}$ given by

$$\text{mse}\left(\hat{\theta}_{\text{RE},i}\right) = g_{1i}\left(\hat{A}_{\text{RE}}\right) + g_{2i}\left(\hat{A}_{\text{RE}}\right) + 2g_{3i}\left(\hat{A}_{\text{RE}}\right). \tag{3.2}$$

The MSE estimator (3.2) is second-order unbiased in the sense that

$$E\left\{\text{mse}\left(\hat{\theta}_{\text{RE},i}\right)\right\} = \text{MSE}\left(\hat{\theta}_{\text{RE},i}\right) + o(m^{-1}).$$

In the case that $A = 0$, the BLUP $\tilde{\theta}_{\text{RE},i}$ of θ_i becomes the regression-synthetic estimator $\hat{\theta}_{\text{SYN},i} = \mathbf{x}_i'\tilde{\boldsymbol{\beta}}(0)$. But surprisingly, the approximation to the MSE of the EBLUP given in (3.1) can be very different from the MSE of the synthetic estimator. Note that the latter is

$$\text{MSE}\left(\hat{\theta}_{\text{SYN},i}\right) = g_{2i}(0) < g_{2i}(0) + g_{3i}(0),$$

because $g_{3i}(0)$ is strictly positive even for $A = 0$. In fact, in the simple example with only one covariate $(p = 1)$ with constant values $x_i = 1$ and constant sampling variances $D_i = D, i = 1, \ldots, m$, we have $\text{MSE}\left(\hat{\theta}_{\text{SYN},i}\right) = g_{2i}(0) = D/m$ whereas the approximation to the MSE of the EBLUP given in (3.1) with $A = 0$ gives $\text{MSE}\left(\hat{\theta}_{\text{RE},i}\right) \approx g_{2i}(0) + g_{3i}(0) = 3D/m$, three times larger. It turns out that (3.1) is not a good approximation of the MSE of the EBLUP when $A = 0$ and, instead, we should use $\text{MSE}\left(\hat{\theta}_{\text{RE},i}\right) = g_{2i}(0)$. Moreover, since for $A = 0$ this quantity does not depend on any unknown parameter, we can take it also as MSE estimator, i.e., we can take $\text{mse}\left(\hat{\theta}_{\text{RE},i}\right) = g_{2i}(0)$.

In practice, the true value of A is not known but we have the consistent estimator $\hat{A}_{\text{RE}}$. When $\hat{A}_{\text{RE}} = 0$, the EBLUP becomes the regression-synthetic estimator for all areas, that is

$$\hat{\theta}_{\text{RE},i} = \hat{\theta}_{\text{SYN},i} = \mathbf{x}_i'\tilde{\boldsymbol{\beta}}(0), i = 1, \ldots, m.$$

In this case, $g_{1i}\left(\hat{A}_{\text{RE}}\right) = 0$ for all areas and the MSE estimator given in (3.2) reduces to

$$\text{mse}\left(\hat{\theta}_{\text{RE},i}\right) = g_{2i}(0) + 2g_{3i}(0) > g_{2i}(0) = \text{MSE}\left(\hat{\theta}_{\text{SYN},i}\right), i = 1,\ldots,m.$$

Thus, the MSE estimator given in (3.2) can be seriously overestimating the MSE for $\hat{A}_{\text{RE}} = 0$. To reduce the overestimation, we consider a modified MSE estimator of $\hat{\theta}_{\text{RE},i}$ given by

$$\text{mse}_0\left(\hat{\theta}_{\text{RE},i}\right) = \begin{cases} g_{2i} & \text{if } \hat{A}_{\text{RE}} = 0, \\ g_{1i}\left(\hat{A}_{\text{RE}}\right) + g_{2i}\left(\hat{A}_{\text{RE}}\right) + 2g_{3i}\left(\hat{A}_{\text{RE}}\right) & \text{if } \hat{A}_{\text{RE}} > 0, \end{cases} \tag{3.3}$$

where $g_{2i} = g_{2i}(0) = \mathbf{x}_i'\left(\mathbf{X}'\mathbf{D}^{-1}\mathbf{X}\right)^{-1}\mathbf{x}_i, i = 1,\ldots,m.$

In fact, for A close to zero, it may happen that g_{2i} is closer to the true MSE than the full MSE estimator $\text{mse}\left(\hat{\theta}_{\text{RE},i}\right)$, but the question of when is A close enough to zero arises. This question motivates the use of a preliminary testing procedure of $A = 0$ to define alternative MSE estimators of the EBLUP in Section 4.

4 Preliminary test estimators

The estimator of A used in the EBLUP of θ_i introduces uncertainty, which might not be negligible for small m. Indeed, the term g_{3i} in the MSE estimator (3.2) arises due to the estimation of A. However, when the value of A is small enough relative to the sampling variances, this uncertainty could be avoided by using the regression-synthetic estimator $\mathbf{x}_i'\tilde{\boldsymbol{\beta}}(0)$ instead of the EBLUP. Datta et al. (2011) proposed a small area estimator based on a preliminary testing procedure of $H_0 : A = 0$ against $H_1 : A > 0$. When H_0 is not rejected, the regression-synthetic estimator is taken as the estimator of θ_i; otherwise, the usual EBLUP is used. They proposed the test statistic

$$T = \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{PT}}\right)' \mathbf{D}^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{PT}}\right),$$

where $\hat{\boldsymbol{\beta}}_{\text{PT}} = \left(\mathbf{X}'\mathbf{D}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{D}^{-1}\mathbf{y}$ is the WLS estimator of $\boldsymbol{\beta}$ obtained assuming that $H_0 : A = 0$ is true. The test statistic T is distributed as X^2_{m-p} with $m - p$ degrees of freedom under H_0. Then, for a specified significance level α, the PTE of $\boldsymbol{\theta}$ defined by Datta et al. (2011) is given by

$$\hat{\boldsymbol{\theta}}_{\text{PT}} = \left(\hat{\theta}_{\text{PT},1},\ldots,\hat{\theta}_{\text{PT},m}\right)' = \begin{cases} \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{PT}} & \text{if } T \le X^2_{m-p,\alpha}; \\ \hat{\boldsymbol{\theta}}_{\text{RE}} & \text{if } T > X^2_{m-p,\alpha}, \end{cases}$$

where $X^2_{m-p,\alpha}$ is the upper α-point of X^2_{m-p}. The PTE is especially designed to handle cases with a modest number of small areas, say $m = 15$.

Here we propose to use the PT procedure for the estimation of MSE of the EBLUP, by considering only the MSE of the synthetic estimator g_{2i} whenever the null hypothesis is not rejected and the full MSE estimate otherwise. But observe that the test statistic T in the PT procedure does not depend on the estimator of A. This means that, even when H_0 is rejected, it may happen that $\hat{A}_{\text{RE}} = 0$. Thus, here we define the PT estimator of the MSE of the EBLUP $\hat{\theta}_{\text{RE},i}$ as

$$\text{mse}_{\text{PT}}\left(\hat{\theta}_{\text{RE},i}\right) = \begin{cases} g_{2i} & \text{if } T \leq X^2_{m-p,\alpha} \quad \text{or} \quad \hat{A}_{\text{RE}} = 0, \\ g_{1i}\left(\hat{A}_{\text{RE}}\right) + g_{2i}\left(\hat{A}_{\text{RE}}\right) + 2g_{3i}\left(\hat{A}_{\text{RE}}\right) & \text{if } T > X^2_{m-p,\alpha} \quad \text{and} \quad \hat{A}_{\text{RE}} > 0. \end{cases} \tag{4.1}$$

5 Adjusted maximum likelihood

The estimation methods for A described in Section 2 might produce zero estimates. In this case, the EBLUPs will give zero weight to the direct estimators in all areas, regardless of the efficiency of the direct estimator in each area. On the other hand, survey sampling practitioners often prefer to give always a strictly positive weight to direct estimators because they are based on the area-specific unit level data for the variable of interest without the assumption of any regression model. For this situation, Li and Lahiri (2010) proposed the AML estimator that delivers a strictly positive estimator of A. This estimator, denoted here $\hat{A}_{\text{AML}}$, is obtained by maximizing the adjusted likelihood defined as

$$L_{\text{AML}}(A) = A \times L_p(A).$$

The EBLUP given in (2.6) with $\hat{A} = \hat{A}_{\text{AML}}$ will be denoted hereafter as $\hat{\boldsymbol{\theta}}_{\text{AML}} = \left(\hat{\theta}_{\text{AML},1}, \ldots, \hat{\theta}_{\text{AML},m}\right)'$. Note that $\hat{\boldsymbol{\theta}}_{\text{AML}}$ assigns strictly positive weights to direct estimators.

Li and Lahiri (2010) proposed a second order unbiased MSE estimator of $\hat{\theta}_{\text{AML},i}$ given by

$$\begin{aligned} \text{mse}\left(\hat{\theta}_{\text{AML},i}\right) &= g_{1i}\left(\hat{A}_{\text{AML}}\right) + g_{2i}\left(\hat{A}_{\text{AML}}\right) + 2g_{3i}\left(\hat{A}_{\text{AML}}\right) \\ &- B_i^2\left(\hat{A}_{\text{AML}}\right) b_{\text{AML}}\left(\hat{A}_{\text{AML}}\right), \end{aligned} \tag{5.1}$$

where $b_{\text{AML}}(A)$ is the bias of $\hat{A}_{\text{AML}}$ and it is given by

$$b_{\text{AML}}(A) = \frac{\text{trace}\left\{\mathbf{P}(A) - \boldsymbol{\Sigma}^{-1}(A)\right\} + 2/A}{\text{trace}\left\{\boldsymbol{\Sigma}^{-2}(A)\right\}}.$$

6 Combined estimators

The strictly positive AML estimator of A has typically a larger bias than ML or REML estimators for A small relative to the D_i's. Thus, if we still wish to obtain a small area estimator that attaches a strictly positive weight to the direct estimator, to reduce the mentioned bias it will be better to use the AML estimator only when strictly necessary; that is, either when data does not provide enough evidence against $A = 0$ or when the resulting REML estimator of A is zero. This section introduces two small area estimators of $\boldsymbol{\theta}$ that give a strictly positive weight to the direct estimator, which are obtained as a combination of the EBLUP based on the AML method and the EBLUP based on REML estimation.

In the first combined proposal, the AML method is used to estimate A when the preliminary test does not reject the null hypothesis and in the second combined proposal, when the REML estimate is non positive. Specifically, the first combined estimator, called hereafter PT-AML, is defined by

$$\hat{\boldsymbol{\theta}}_{\text{PTAML}} = \begin{cases} \hat{\boldsymbol{\theta}}_{\text{AML}} & \text{if } T \leq X^2_{m-p,\alpha} \quad \text{or} \quad \hat{A}_{\text{RE}} = 0, \\ \hat{\boldsymbol{\theta}}_{\text{RE}} & \text{if } T > X^2_{m-p,\alpha} \quad \text{and} \quad \hat{A}_{\text{RE}} > 0. \end{cases} \tag{6.1}$$

The second combined estimator, called REML-AML, is given by

$$\hat{\boldsymbol{\theta}}_{\text{REAML}} = \begin{cases} \hat{\boldsymbol{\theta}}_{\text{AML}} & \text{if } \hat{A}_{\text{RE}} = 0, \\ \hat{\boldsymbol{\theta}}_{\text{RE}} & \text{if } \hat{A}_{\text{RE}} > 0, \end{cases} \tag{6.2}$$

see Rubin-Bleuer and Yu (2013). For the estimation of MSE of $\hat{\boldsymbol{\theta}}_{\text{REAML}}$, these authors proposed

$$\text{mse}\left(\hat{\theta}_{\text{REAML},i}\right) = \begin{cases} \text{mse}\left(\hat{\theta}_{\text{AML},i}\right) & \text{if } \hat{A}_{\text{RE}} = 0, \\ \text{mse}\left(\hat{\theta}_{\text{RE},i}\right) & \text{if } \hat{A}_{\text{RE}} > 0. \end{cases} \tag{6.3}$$

Using $\text{mse}\left(\hat{\theta}_{\text{AML},i}\right)$ when $\hat{A}_{\text{RE}} = 0$ leads to substantial overestimation if the true value of A is small because $\hat{\theta}_{\text{AML},i}$ will be closer to the regression-synthetic estimator. Hence, we propose the alternative MSE estimator

$$\text{mse}_0\left(\hat{\theta}_{\text{REAML},i}\right) = \begin{cases} g_{2i} & \text{if } \hat{A}_{\text{RE}} = 0, \\ \text{mse}\left(\hat{\theta}_{\text{RE},i}\right) & \text{if } \hat{A}_{\text{RE}} > 0. \end{cases} \tag{6.4}$$

Again, since for small A, $\text{mse}\left(\hat{\theta}_{\text{RE},i}\right)$ might still be overestimating the true MSE of $\hat{\theta}_{\text{REAML},i}$, we consider also the following PT estimator

$$\text{mse}_{\text{PT}}\left(\hat{\theta}_{\text{REAML},i}\right) = \begin{cases} g_{2i} & \text{if } T \leq X^2_{m-p,\alpha} \quad \text{or} \quad \hat{A}_{\text{RE}} = 0, \\ \text{mse}\left(\hat{\theta}_{\text{RE},i}\right) & \text{if } T > X^2_{m-p,\alpha} \quad \text{and} \quad \hat{A}_{\text{RE}} > 0. \end{cases} \tag{6.5}$$

7 Simulation experiments

A simulation study was designed with the following purposes in mind:

(a) To study the properties, in terms of bias and MSE, of the PT estimators as α varies for fixed A and as A varies for fixed α. We would like to see which values of α are adequate for a given A.

(b) To compare the PTEs with the EBLUPs based on REML and with the EBLUPs based on AML.

(c) To study the performance of the proposed MSE estimators in terms of relative bias and also in terms of coverage and length of prediction intervals.

(d) To compare the three introduced small area estimators that give strictly positive weight to the direct estimator for all areas, namely EBLUP based on AML, PT-AML and REML-AML estimators.

To accomplish the above goals, data were generated from the Fay-Herriot model given by (2.1)-(2.2) with a constant mean, that is, with $p = 1, \boldsymbol{\beta} = \mu$ and $\mathbf{x}_i = 1, i = 1, \ldots, m$. We let $\mu = 0$ without loss of

generality, number of areas $m = 15$ and $D_i = 1, i = 1, \ldots, m$. The simulation study was repeated for increasing values of the model variance, $A \in \{0.01, 0.02, 0.05, 0.1, 0.2, 1\}$, and also for six significance levels of the test of $H_0 : A = 0$ against $H_0 : A > 0$, namely $\alpha = \{0.05, 0.1, 0.2, 0.3, 0.4, 0.5\}$. For each combination of A and α, the following steps were performed for each simulation run $\ell = 1, \ldots, L$ with $L = 10{,}000$ runs:

1. Generate data from the assumed model with constant zero mean; i.e.,

$$\begin{aligned} \theta_i^{(\ell)} &= v_i^{(\ell)}, \quad v_i^{(\ell)} \overset{\text{ind}}{\sim} N(0, A), \\ y_i^{(\ell)} &= \theta_i^{(\ell)} + e_i^{(\ell)}, \quad e_i^{(\ell)} \overset{\text{ind}}{\sim} N(0, D_i), \quad i = 1, \ldots, m. \end{aligned}$$

2. Calculate the following estimators of $\boldsymbol{\theta}$: the EBLUP based on REML estimation of $A, \hat{\boldsymbol{\theta}}_{\text{RE}}^{(\ell)}$, the PT estimate $\hat{\boldsymbol{\theta}}_{\text{PT}}^{(\ell)}$, the EBLUP based on AML estimation of $A, \hat{\boldsymbol{\theta}}_{\text{AML}}^{(\ell)}$, the combined PT-AML estimate $\hat{\boldsymbol{\theta}}_{\text{PTAML}}^{(\ell)}$ and the REML-AML estimate $\hat{\boldsymbol{\theta}}_{\text{REAML}}^{(\ell)}$.
3. For each area $i = 1, \ldots, m$, calculate: the three estimates of the MSE of the EBLUP $\hat{\theta}_{\text{RE},i}$ given in (3.2), (3.3) and (4.1), denoted respectively by $\text{mse}^{(\ell)}\left(\hat{\theta}_{\text{RE},i}\right), \text{mse}_0^{(\ell)}\left(\hat{\theta}_{\text{RE},i}\right)$ and $\text{mse}_{\text{PT}}^{(\ell)}\left(\hat{\theta}_{\text{RE},i}\right)$, and the three estimates (6.3), (6.4) and (6.5) of the MSE of the combined small area estimator $\hat{\theta}_{\text{REAML},i}$, denoted $\text{mse}^{(\ell)}\left(\hat{\theta}_{\text{REAML},i}\right), \text{mse}_0^{(\ell)}\left(\hat{\theta}_{\text{REAML},i}\right)$ and $\text{mse}_{\text{PT}}^{(\ell)}\left(\hat{\theta}_{\text{REAML},i}\right)$ respectively.
4. For each area $i = 1, \ldots, m$, obtain the normality-based $1 - \alpha$ prediction intervals for the small area mean θ_i based on the three considered MSE estimators of the EBLUP:

$$\begin{aligned} \text{CI}_i^{(\ell)} &= \hat{\theta}_{\text{RE},i}^{(\ell)} \mp Z_{\alpha/2}\sqrt{\text{mse}^{(\ell)}\left(\hat{\theta}_{\text{RE},i}\right)}, \\ \text{CI}_{0,i}^{(\ell)} &= \hat{\theta}_{\text{RE},i}^{(\ell)} \mp Z_{\alpha/2}\sqrt{\text{mse}_0^{(\ell)}\left(\hat{\theta}_{\text{RE},i}\right)}, \\ \text{CI}_{\text{PT},i}^{(\ell)} &= \hat{\theta}_{\text{RE},i}^{(\ell)} \mp Z_{\alpha/2}\sqrt{\text{mse}_{\text{PT}}^{(\ell)}\left(\hat{\theta}_{\text{RE},i}\right)}, \end{aligned}$$

   where $Z_{\alpha/2}$ is the upper $\alpha/2$- point of a standard normal distribution.
5. Repeat Steps 1-4 for $\ell = 1, \ldots, L$, for $L = 10{,}000$. Then, for each small area estimator $\hat{\theta}_i \in \left\{\hat{\theta}_{\text{RE},i}, \hat{\theta}_{\text{PT},i}, \hat{\theta}_{\text{AML},i}, \hat{\theta}_{\text{PTAML},i}, \hat{\theta}_{\text{REAML},i}\right\}, i = 1, \ldots, m$, compute its empirical bias and MSE as

$$B\left(\hat{\theta}_i\right) = \frac{1}{L}\sum_{\ell=1}^{L}\left(\hat{\theta}_i^{(\ell)} - \theta_i^{(\ell)}\right), \quad \text{MSE}\left(\hat{\theta}_i\right) = \frac{1}{L}\sum_{\ell=1}^{L}\left(\hat{\theta}_i^{(\ell)} - \theta_i^{(\ell)}\right)^2.$$

   Then obtain the average over areas of absolute biases and MSEs as

$$\overline{\text{AB}}\left(\hat{\boldsymbol{\theta}}\right) = \frac{1}{m}\sum_{i=1}^{m}\left|B\left(\hat{\theta}_i\right)\right|, \quad \overline{\text{AMSE}}\left(\hat{\boldsymbol{\theta}}\right) = \frac{1}{m}\sum_{i=1}^{m}\text{MSE}\left(\hat{\theta}_i\right).$$

6. Calculate the relative bias of each MSE estimator, $\text{mse}(\hat{\theta}_i)$, as follows

$$\text{RB}\{\text{mse}(\hat{\theta}_i)\} = \left\{\frac{1}{L}\sum_{\ell=1}^{L}\text{mse}^{(\ell)}(\hat{\theta}_i) - \text{MSE}(\hat{\theta}_i)\right\} \Big/ \text{MSE}(\hat{\theta}_i).$$

Calculate the average over areas of the absolute relative biases as

$$\overline{\text{ARB}}\{\text{mse}(\hat{\boldsymbol{\theta}})\} = \frac{1}{m}\sum_{i=1}^{m}\left|\text{RB}\{\text{mse}(\hat{\theta}_i)\}\right|.$$

7. For each type of prediction interval $\text{CI}_i^{(\ell)} = (L_i^{(\ell)}, U_i^{(\ell)})$, for $\text{CI}_i^{(\ell)} \in \{\text{CI}_i^{(\ell)}, \text{CI}_{0,i}^{(\ell)}, \text{CI}_{\text{PT},i}^{(\ell)}\}$ given in Step 4, calculate the empirical coverage rate (CR) and the average length (AL) as

$$\text{CR}(\text{CI}_i) = \frac{\#\{\theta_i^{(\ell)} \in \text{CI}_i^{(\ell)}\}}{L}, \quad \text{AL}(\text{CI}_i) = \frac{1}{L}\sum_{\ell=1}^{L}(U_i^{(\ell)} - L_i^{(\ell)}).$$

Finally, average over areas the coverage rates and average lengths, as

$$\overline{\text{CR}}(\text{CI}) = \frac{1}{m}\sum_{i=1}^{m}\text{CR}(\text{CI}_i), \quad \overline{\text{AL}}(\text{CI}) = \frac{1}{m}\sum_{i=1}^{m}\text{AL}(\text{CI}_i).$$

Figures 7.1 and 7.2 plot the average MSEs of the PTEs for each $A \in \{0.05, 0.1, 0.2\}$, together with the average MSE of the EBLUPs based on REML and AML, against the significance level α. Note that when A is small, for large α the PT procedure is rejecting H_0 more often and therefore the PTE becomes more often the usual EBLUP, whereas for small α the PT procedure rejects H_0 less often and the regression-synthetic estimator is then more often used. In contrast, for a large value of A, the PTE becomes the EBLUP more frequently regardless of α. The absolute biases of the estimators are not shown here because they are roughly the same for all the PTEs across α values. The reason for this is that when the model holds, both components of the PTE, the synthetic estimator and the EBLUP, are unbiased for the target parameter. Note that the synthetic estimator is unbiased even when $A > 0$. The first conclusion arising from Figures 7.1 and 7.2 is that the MSE of the PTE is practically constant across $\alpha \geq 0.1$. See also that the average MSE of the PTE for a given α increases with A because the PTE reduces to the EBLUP more often as A increases and the MSE of the EBLUP increases with A. Observe also that the PTE and the EBLUP based on REML perform very similarly for $\alpha \geq 0.2$. However, for $\alpha < 0.2$, the PTE becomes more efficient than the EBLUP as soon as A moves close to the null hypothesis $(A < 0.1)$, which agrees with the remark of Datta et al. (2011).

Turning to the EBLUP based on AML, Figures 7.1 and 7.2 show that its average MSE is significantly larger than that of the other two estimators, but the differences with the other ones decrease as A increases. This is due to bias of the AML estimator of A for small A. We shall study later the combined small area estimators PT-AML and REML-AML, which use the EBLUP based on AML only when null hypothesis is not rejected or when the realized estimate of A is zero.

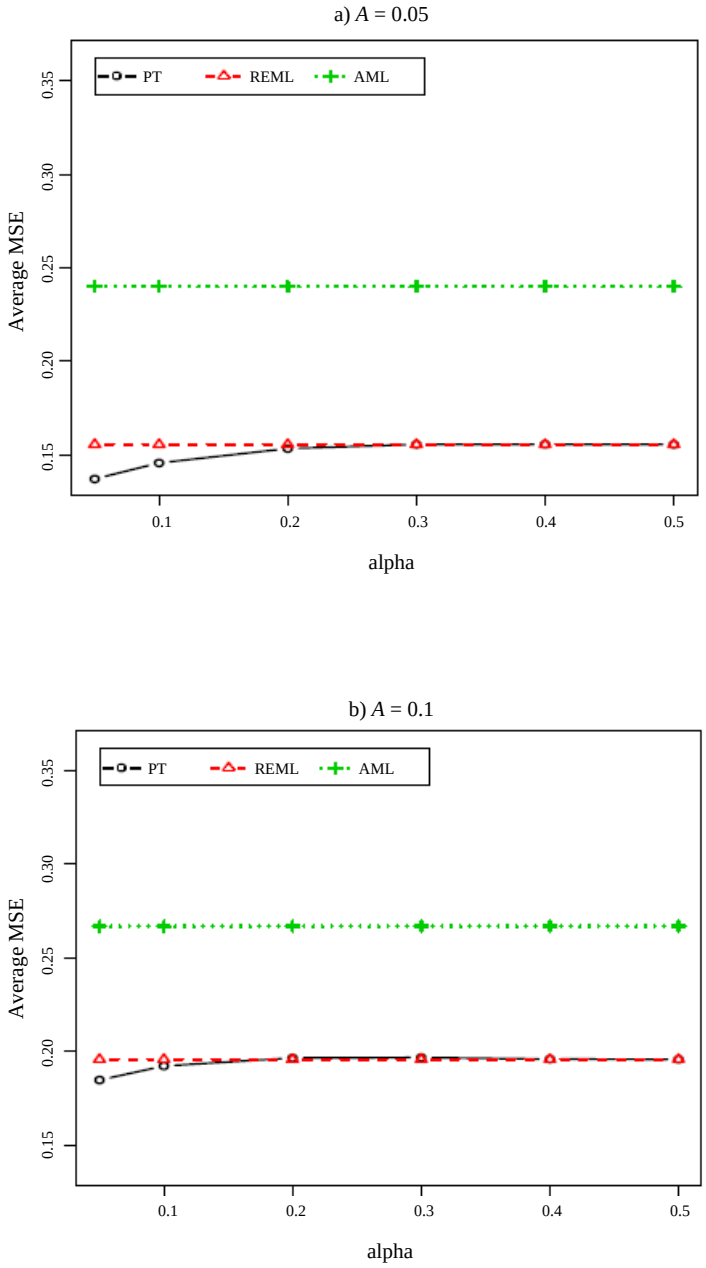


Figure 7.1 Average MSEs of PTE, EBLUP based on REML and EBLUP based on AML against α, for a) $A = 0.05$ and b) $A = 0.1$.

Datta et al. (2011, page 366) recommended $\alpha \geq 0.2$ for the PTE. Moreover, the literature on PT estimation for fixed effects models suggests that a good choice of α in terms of bias and MSE is $\alpha = 0.2$ (Bancroft 1944; Han and Bancroft 1968). But the above results suggest that for $\alpha \geq 0.2$, the PTE is practically the same as the EBLUP and therefore one might choose to always use the EBLUP.

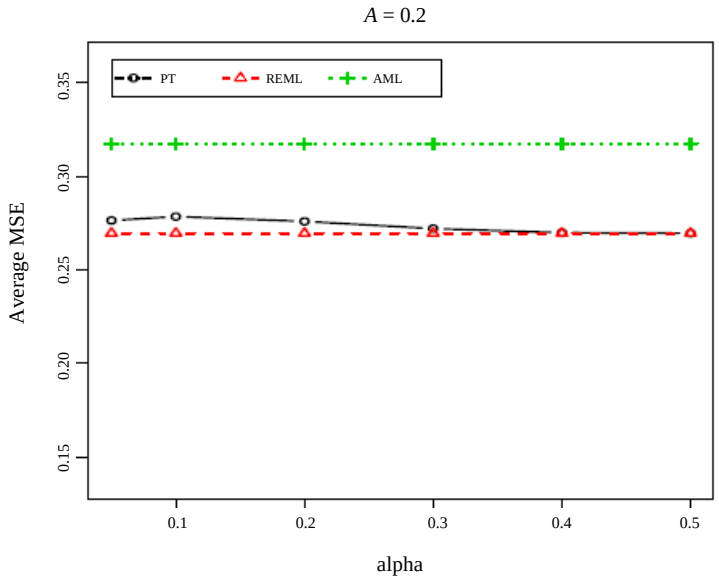


Figure 7.2 Average MSEs of PTE, EBLUP based on REML and EBLUP based on AML against α, for $A = 0.2$.

Now we study the properties of the PT for MSE estimation in terms of α. Figure 7.3 plots the average absolute relative bias of the MSE estimators $\text{mse}_{\text{PT}}\left(\hat{\theta}_{\text{RE},i}\right)$ labelled PT, against the significance level α, for each value $A \in \{0.05, 0.1, 0.2, 1\}$. When α is taken very small $\alpha < 0.1$, the null hypothesis $H_0 : A = 0$ is less often rejected and $\text{mse}_{\text{PT}}\left(\hat{\theta}_{\text{RE},i}\right)$ becomes often g_{2i}, which leads to underestimation. For α large $(\alpha > 0.2)$, the null hypothesis is more often rejected and $\text{mse}_{\text{PT}}\left(\hat{\theta}_{\text{RE},i}\right)$ becomes the usual MSE estimator of the EBLUP, which severely overestimates the true MSE for small A. The value $\alpha = 0.2$ appears to be a good compromise choice, with an average absolute relative bias around 10% for $A \geq 0.1$ and 20% for $A = 0.05$.

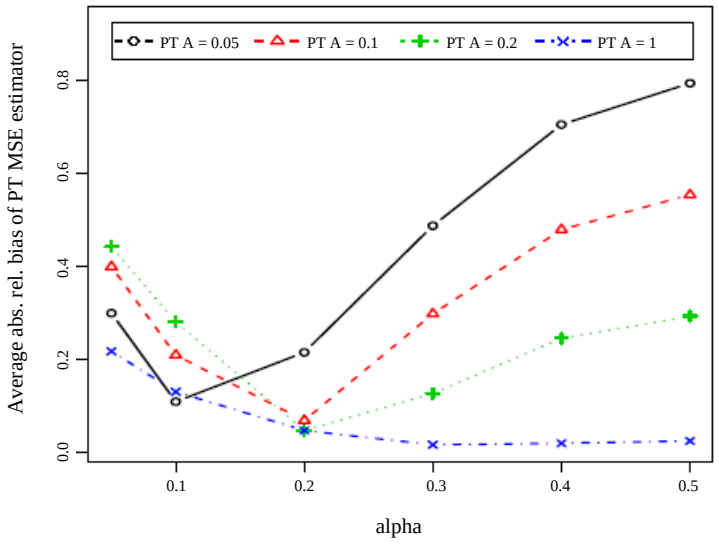


Figure 7.3 Average over areas of absolute relative biases of the MSE estimator $\text{mse}_{\text{PT}}\left(\hat{\theta}_{\text{RE},i}\right)$, labelled PT, for $A \in \{0.05, 0.1, 0.2, 1\}$ against significance level α.

The above results suggest that $\alpha = 0.2$ is a good choice when using the PT procedure to estimate the MSE of the usual EBLUP. This has been more thoroughly studied by looking at the (signed) relative biases of $\text{mse}_{\text{PT}}\left(\hat{\theta}_{\text{RE},i}\right)$ for each area. These results are plotted in Figures 7.4 and 7.5 with four plots, one for each value of $A \in \{0.05, 0.1, 0.2, 1\}$. The figures appearing in the legends of these plots are the significance levels α for the PT MSE estimator $\text{mse}_{\text{PT}}\left(\hat{\theta}_{\text{RE},i}\right)$. These plots confirm our previous observations: the MSE estimator based on the PT, $\text{mse}_{\text{PT}}\left(\hat{\theta}_{\text{RE},i}\right)$, underestimates $\text{MSE}\left(\hat{\theta}_{\text{RE},i}\right)$ for small α and overestimates for large α. It turns out that $\text{mse}_{\text{PT}}\left(\hat{\theta}_{\text{RE},i}\right)$ with $\alpha = 0.2$ is a good candidate for all values of A.

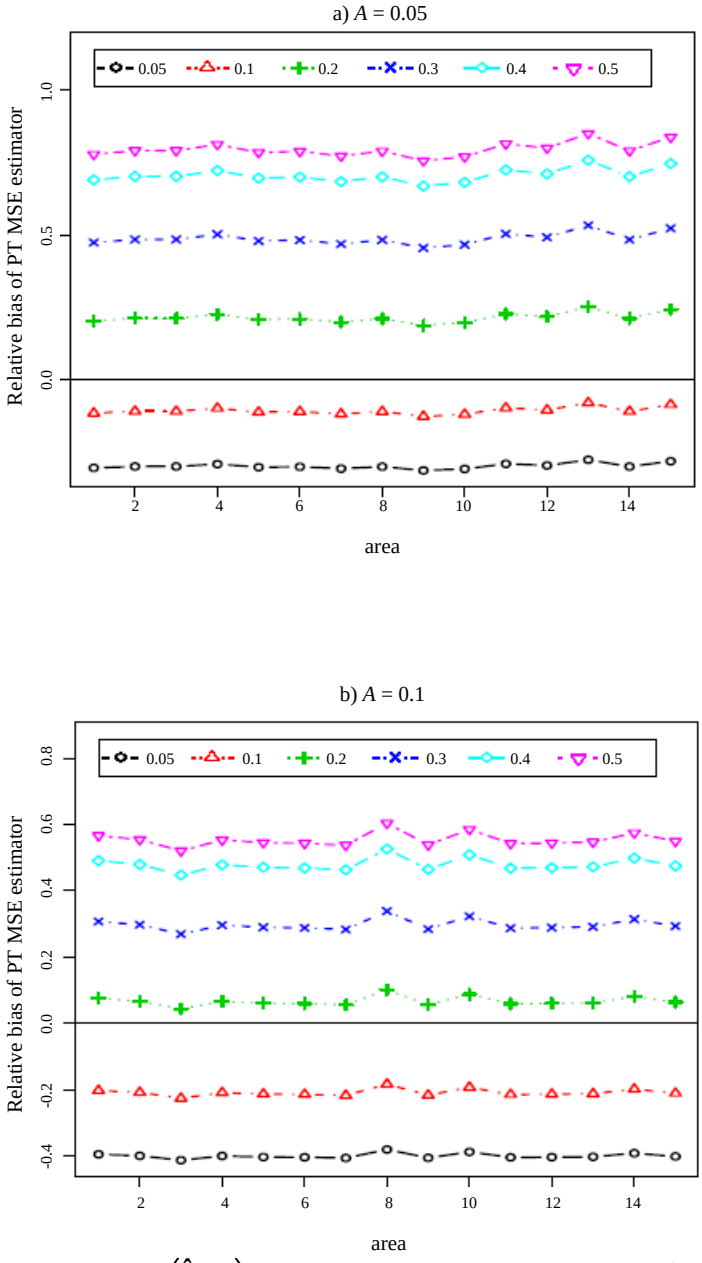


Figure 7.4 Relative biases of $\text{mse}_{\text{PT}}\left(\hat{\theta}_{\text{RE},i}\right)$, for each significance level $\alpha \in \{0.05, 0.1, 0.2, 0.3, 0.4, 0.5\}$, against area i, for a) $A = 0.05$ and b) $A = 0.1$.

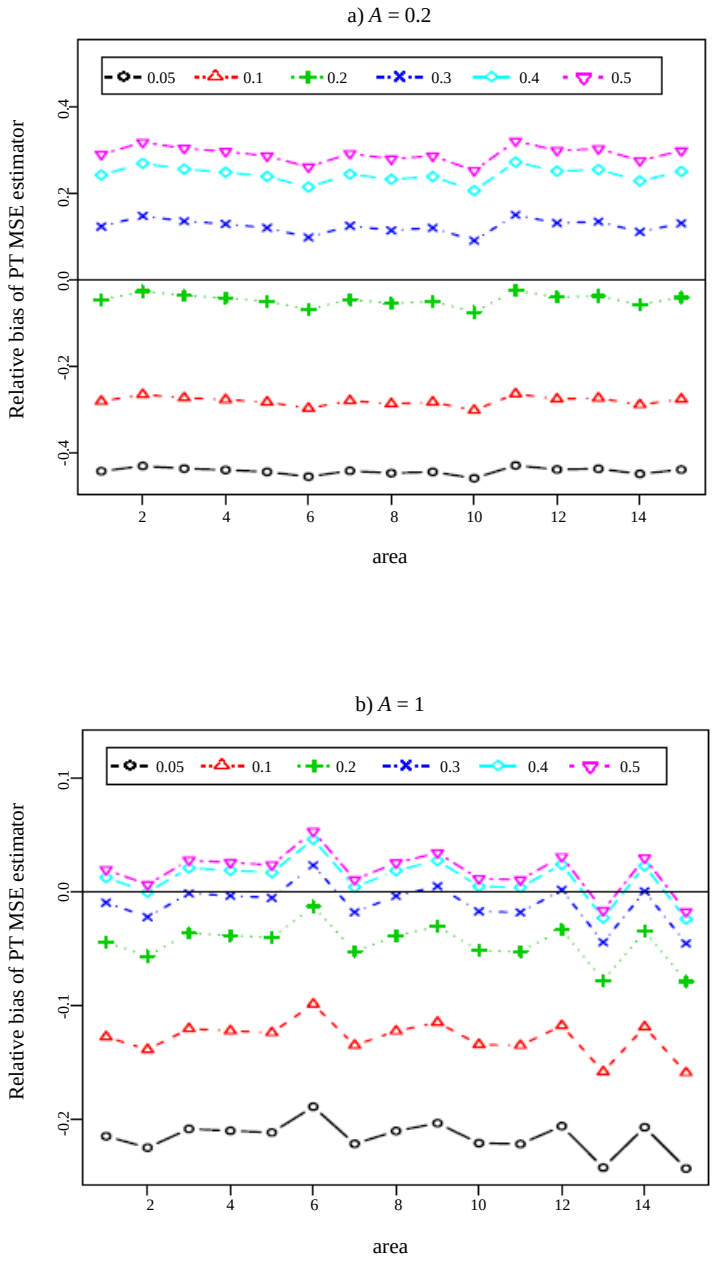


Figure 7.5 Relative biases of $\text{mse}_{\text{PT}}\left(\hat{\theta}_{\text{RE},i}\right)$, for each significance level $\alpha \in \{0.05, 0.1, 0.2, 0.3, 0.4, 0.5\}$, against area i, for a) $A = 0.2$ and b) $A = 1$.

Let us now compare $\text{mse}_{\text{PT}}\left(\hat{\theta}_{\text{RE},i}\right)$ for the selected significance level $\alpha = 0.2$ with the other two MSE estimators $\text{mse}_{0}\left(\hat{\theta}_{\text{RE},i}\right)$ and $\text{mse}\left(\hat{\theta}_{\text{RE},i}\right)$ given by (3.3) and (3.2) respectively. Figure 7.6 plots the average absolute relative biases of the three MSE estimators, labelled respectively PT, REML0 and REML. We note that $\text{mse}_{0}\left(\hat{\theta}_{\text{RE},i}\right)$ performs better than $\text{mse}\left(\hat{\theta}_{\text{RE},i}\right)$ for all areas, but still $\text{mse}_{\text{PT}}\left(\hat{\theta}_{\text{RE},i}\right)$ is better than $\text{mse}_{0}\left(\hat{\theta}_{\text{RE},i}\right)$ for all considered values of A except for $A = 1$, where the differences between the three estimators are negligible. The differences decrease as A increases, but observe that the usual MSE estimator, $\text{mse}\left(\hat{\theta}_{\text{RE},i}\right)$, can be severely biased for small A, with an average absolute relative

bias over 50% for $A < 0.2$ and exponentially growing as A tends to zero. The conclusion is that, when H_0 is not rejected, even if the realized estimate of A is positive, it seems better to omit the g_{3i} term in the MSE estimator and consider only g_{2i}.

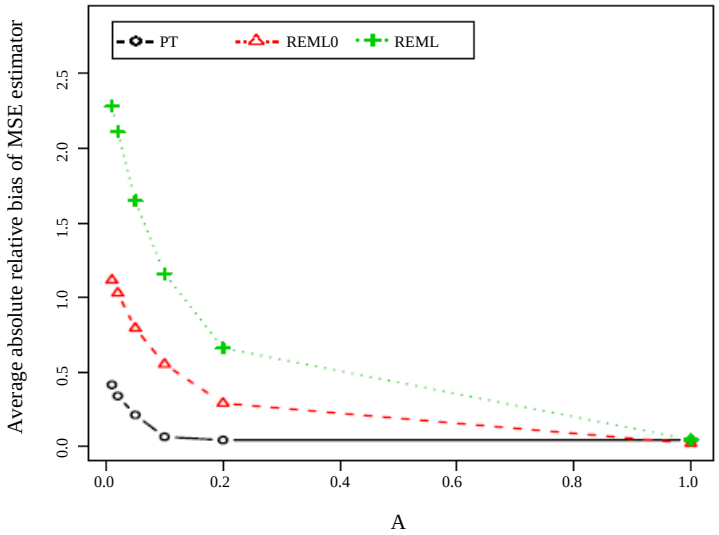


Figure 7.6 Average over areas of absolute relative biases of MSE estimators $\text{mse}_{\text{PT}}\left(\hat{\theta}_{\text{RE},i}\right)$ with $\alpha = 0.2$, labelled PT, $\text{mse}\left(\hat{\theta}_{\text{RE},i}\right)$ labelled REML and $\text{mse}_0\left(\hat{\theta}_{\text{RE},i}\right)$ labelled REML0, against A.

We now turn to the small area estimators that attach strictly positive weight to the direct estimator for all areas: EBLUP based on AML, $\hat{\theta}_{\text{AML}}$, and the two combined estimators, PT-AML given in (6.1), and REML-AML given in (6.2). Average MSEs are plotted in Figure 7.7 for these three estimators. In this plot, $\hat{\theta}_{\text{AML}}$ seems to be a little less efficient, followed by PT-AML. The combined estimator REML-AML seems to perform slightly better than its two counterparts for small A, although for $A \geq 0.2$ the PT-AML estimator is very close to it. For MSE estimation, we focus on REML-AML because of its better performance.

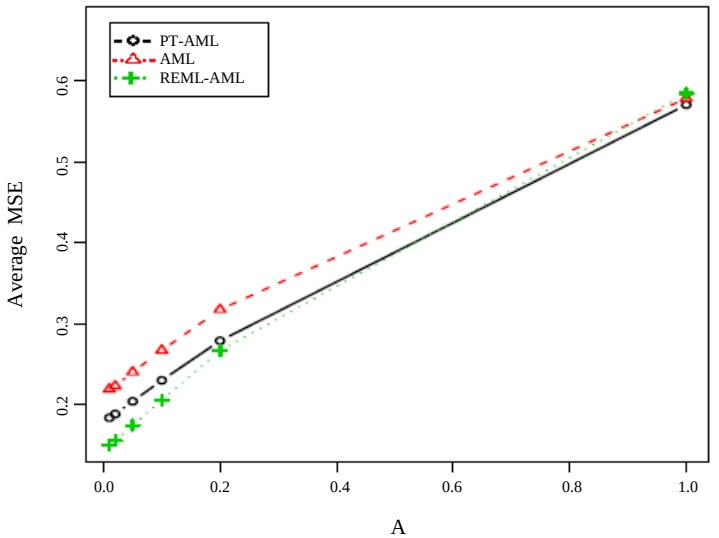


Figure 7.7 Average over areas of MSEs of PT-AML estimator with $\alpha = 0.2$, EBLUP based on AML and REML-AML estimator against A.

For the combined estimator REML-AML, Figure 7.8 shows that the MSE estimator based on the PT, $\text{mse}_{\text{PT}}\left(\hat{\theta}_{\text{REAML},i}\right)$, which uses only g_{2i} whenever $\hat{A}_{\text{RE}} = 0$ or the null hypothesis is not rejected, has average absolute relative bias less than 10% for $A \geq 0.1$ and it is smaller than the corresponding values for $\text{mse}\left(\hat{\theta}_{\text{REAML},i}\right)$ and $\text{mse}_0\left(\hat{\theta}_{\text{REAML},i}\right)$, especially for $A \leq 0.4$.

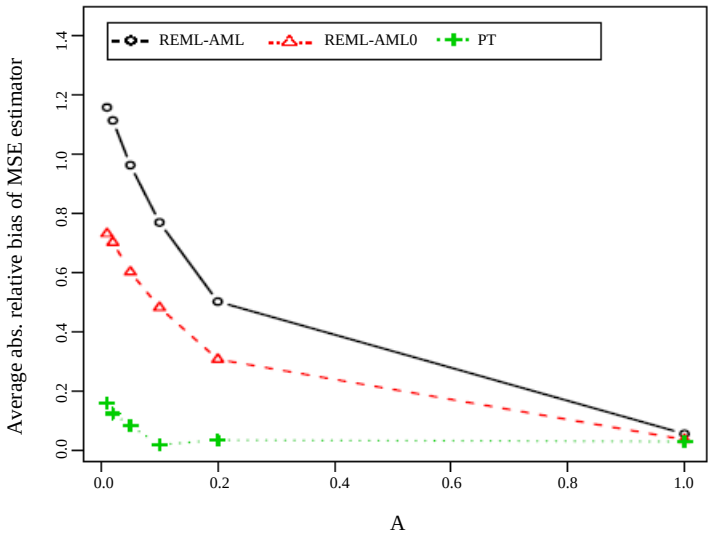


Figure 7.8 Average over areas of absolute relative biases of the MSE estimators $\text{mse}\left(\hat{\theta}_{\text{REAML},i}\right)$, $\text{mse}_0\left(\hat{\theta}_{\text{REAML},i}\right)$ and $\text{mse}_{\text{PT}}\left(\hat{\theta}_{\text{REAML},i}\right)$, labelled respectively REML-AML, REML-AML0 and PT, against *A*.

Finally, we analyze the average over areas of coverage rates and average lengths of normality-based prediction intervals for the small area mean θ_i using the EBLUP based on REML as point estimate and the three different MSE estimators of the EBLUP, namely $\text{mse}\left(\hat{\theta}_{\text{RE},i}\right)$, $\text{mse}_0\left(\hat{\theta}_{\text{RE},i}\right)$ and $\text{mse}_{\text{PT}}\left(\hat{\theta}_{\text{RE},i}\right)$. Figure 7.9 shows the coverage rates of these three types of intervals, where the MSE estimators based on the PT procedure were obtained taking $\alpha = 0.2, 0.3$. It seems that the good relative bias properties of the MSE estimator based on the PT, $\text{mse}_{\text{PT}}\left(\hat{\theta}_{\text{RE},i}\right)$, for small A cannot be extrapolated to coverage based on normal prediction intervals, showing undercoverage especially for $A = 0.2$. In this case, taking a larger significance level $\alpha = 0.3$ reduces a little the undercoverage of the prediction intervals obtained using $\text{mse}_{\text{PT}}\left(\hat{\theta}_{\text{RE},i}\right)$. Still, the coverage rates of $\text{mse}_0\left(\hat{\theta}_{\text{RE},i}\right)$ are better for all values of A. As expected, the usual MSE estimator $\text{mse}\left(\hat{\theta}_{\text{RE},i}\right)$ provides overcoverage for small values of A, which is due to the severe overestimation of the MSE. On the other hand, the intervals showing undercoverage also lead to shorter prediction intervals as shown by Figure 7.10.

It is worthwhile to mention that the construction of prediction intervals for θ_i based on the Fay-Herriot model with accurate coverage rates is not an obvious task. Several papers have appeared in the literature for this problem. For example, Chatterjee, Lahiri and Li (2008) proposed prediction intervals with second order correct coverage rate using only the g_{1i} term as MSE estimate and applying a bootstrap procedure to find the calibrated quantiles. Diao, Smith, Datta, Maiti and Opsomer (2014) have recently

obtained prediction intervals with second order correct coverage rate avoiding the use of resampling procedures and using the full MSE estimator. Obtaining prediction intervals with accurate coverage using other MSE estimates is still a challenge and it is out of scope of this paper.

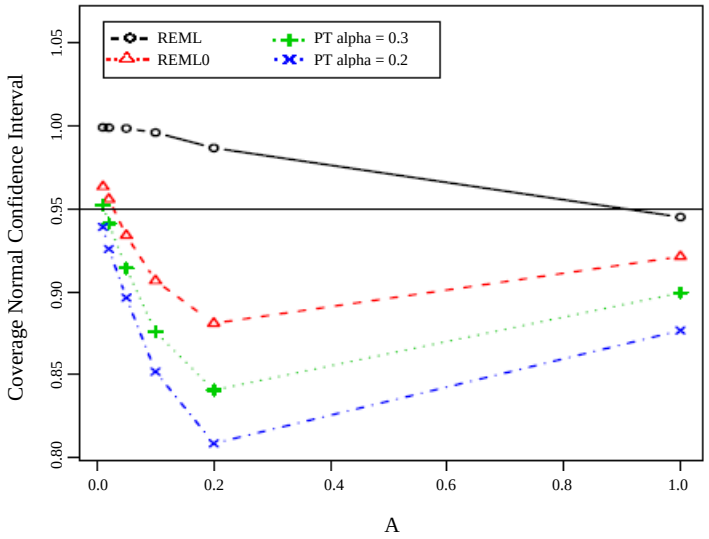


Figure 7.9 Average over areas of coverage rates of normality-based prediction intervals for θ_i using the MSE estimators $\text{mse}\left(\hat{\theta}_{\text{RE},i}\right), \text{mse}_0\left(\hat{\theta}_{\text{RE},i}\right)$ and $\text{mse}_{\text{PT}}\left(\hat{\theta}_{\text{RE},i}\right)$ with $\alpha = 0.2, 0.3$, labelled respectively REML, REML0 and PT, against *A*.

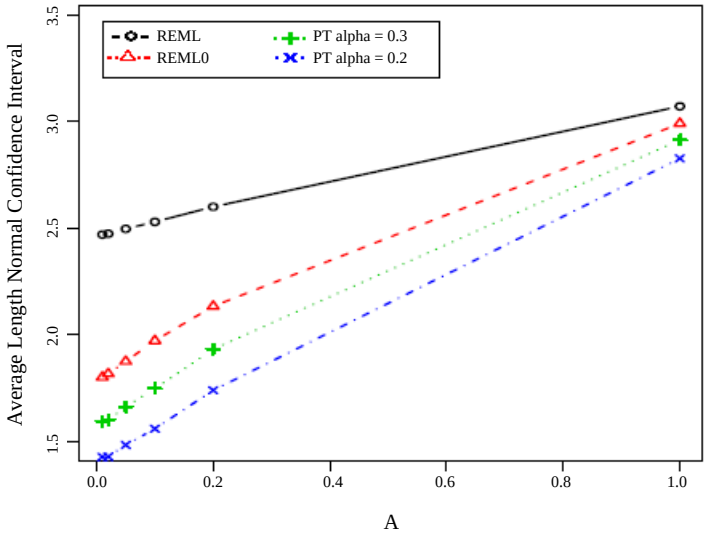


Figure 7.10 Average over areas of average lengths of normality-based intervals for θ_i using the MSE estimators $\text{mse}\left(\hat{\theta}_{\text{RE},i}\right), \text{mse}_0\left(\hat{\theta}_{\text{RE},i}\right)$ and $\text{mse}_{\text{PT}}\left(\hat{\theta}_{\text{RE},i}\right)$ with $\alpha = 0.2, 0.3$, labelled respectively REML, REML0 and PT, against *A*.

This simulation study described above was repeated for several patterns of unequal sampling variances D_i. Although results are not reported here, conclusions are very similar as long as the variance pattern is not extremely uneven.

8 Conclusions

The following major conclusions may be drawn from the results of our simulation study on the estimation of small area means, based on the Fay-Herriot area-level model when the number of areas is modest in size (say $m = 15$) : 1) Under the Fay-Herriot model with a value of random effects variance, A, clearly away from zero, the PTE does not seem to noticeably improve efficiency relative to the usual EBLUP unless the significance level is taken small ($\alpha \leq 0.1$ in our simulation study). 2) Our simulation results indicate that using the PT procedure with a moderate α, in particular $\alpha = 0.2$, to estimate the MSE of the usual EBLUP leads to a reduction in bias as compared with the usual MSE estimator. Hence, we recommend the use of $\text{mse}_{\text{PT}}\left(\hat{\theta}_{\text{RE},i}\right)$, given by (4.1), to estimate the MSE of the EBLUP. 3) Among the estimators that attach a strictly positive weight to the direct estimator for all areas, we recommend the combined estimator REML-AML given by (6.2), because it achieves slightly higher efficiency than the EBLUP based on AML and the PT-AML given by (6.1). 4) For estimating the MSE of the recommended REML-AML estimator, the estimator $\text{mse}_{\text{PT}}\left(\hat{\theta}_{\text{REAML},i}\right)$ given by (6.5) performs better than the alternative ones. 5) Our results on prediction intervals, based on normal theory, indicate that the good performance of the proposed MSE estimators may not translate to coverage properties of these intervals. Construction of prediction intervals that lead to accurate coverages, using the proposed MSE estimates, appears to be a difficult task.

Smooth alternatives to the preliminary test estimates in the case of location parameters have been proposed in the literature using weighted means of the estimates obtained under the null and alternative hypotheses, with weights depending on the test statistic, see e.g., Saleh (2006). Mean squared error estimates of this kind have not been studied and we leave this subject for further research.

Acknowledgements

We would like to thank the editor for very constructive suggestions. Gauri S. Datta's research was partially supported through the grant H98230-11-1-0208 from the National Security Agency, Isabel Molina's research by grants ref. MTM2009-09473, MTM2012-37077-C02-01 and SEJ2007-64500 from the Spanish *Ministerio de Educación y Ciencia* and J.N.K. Rao's research by the Natural Sciences and Engineering Research Council of Canada.

References

Bancroft, T.A. (1944). On biases in estimation due to the use of preliminary tests of significance. *The Annals of Mathematical Statistics*, 15, 190-204.

Chatterjee, S., Lahiri, P. and Li, H. (2008). Parametric bootstrap approximation to the distribution of EBLUP and related prediction intervals in linear mixed models. *The Annals of Statistics*, 36, 1221-1245.

Das, K., Jiang, J. and Rao, J.N.K. (2004). Mean squared error of empirical predictor. *The Annals of Statistics*, 32, 818-840.

Datta, G.S., and Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica*, 10, 613-627.

Datta, S., Hall, P. and Mandal, A. (2011). Model selection by testing for the presence of small-area effects, and application to area-level data. *Journal of the American Statistical Association*, 106, 362-374.

Diao, L., Smith, D.D., Datta, G.S., Maiti, T. and Opsomer, J.D. (2014). Accurate confidence interval estimation of small area parameters under the Fay-Herriot model. *Scandinavian Journal of Statistics*, to appear.

Fay, R.E., and Herriot, R.A. (1979). Estimation of income from small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

Han, C.-P., and Bancroft, T.A. (1968). On pooling means when variance is unknown. *Journal of the American Statistical Association*, 63, 1333-1342.

Li, H., and Lahiri, P. (2010). An adjusted maximum likelihood method for solving small area estimation problems. *Journal of Multivariate Analysis*, 101, 882-892.

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

Rao, J.N.K. (2003). *Small Area Estimation*. Hoboken, NJ: Wiley.

Rubin-Bleuer, S., and Yu, Y. (2013). A positive variance estimator for the Fay-Herriot small area model. SRID2-12-001E, Statistics Canada.

Saleh, A.K. Md. E. (2006). *Theory of Preliminary Test and Stein-type Estimation with Applications*. New York: John Wiley & Sons, Inc.

Survey Methodology, June 2015
Vol. 41, No. 1, pp. 21-36
Statistics Canada, Catalogue No. 12-001-X

Small area estimation combining information from several sources

Jae-kwang Kim, Seunghwan Park and Seo-young Kim[1]

Abstract

An area-level model approach to combining information from several sources is considered in the context of small area estimation. At each small area, several estimates are computed and linked through a system of structural error models. The best linear unbiased predictor of the small area parameter can be computed by the general least squares method. Parameters in the structural error models are estimated using the theory of measurement error models. Estimation of mean squared errors is also discussed. The proposed method is applied to the real problem of labor force surveys in Korea.

Key Words: Area-level model; Auxiliary information; Measurement error models; Structural error model; Survey integration.

1 Introduction

Combining information from different sources is an important problem in statistics. In survey sampling, combining information from multiple surveys can improve the quality of small area estimates. The source of information can come from a probability sample with direct measurements, from another probability sample with indirect measurements (such as self-reported health status), or from auxiliary area-level information. Many approaches of combining information, such as the multiple-frame and statistical matching methods, require access to individual level data, which is not always feasible in practice.

We consider an area-level model approach to small area estimation when there are several sources of auxiliary information. Pfeffermann (2002) and Rao (2003) provided thorough reviews of methods used in small area estimation. Lohr and Prasad (2003) used multivariate models to combine information from several surveys. Ybarra and Lohr (2008) considered the small area estimation problem when the area-level auxiliary information has measurement errors. Merkouris (2010) discussed the small area estimation by combining information from multiple surveys. Raghunathan, Xie, Schenker, Parsons, Davis, Dodd and Feuer (2007) and Manzi, Spiegelhalter, Turner, Flowers and Thompson (2011) used Bayesian hierarchical models to combine information from multiple surveys for small area estimation. Kim and Rao (2012) considered a design-based approach to combining information from two independent surveys.

To describe the setup, suppose that the finite population consists of H subpopulations, denoted by $U_1, \ldots, U_H$, and that we are interested in estimating the subpopulation totals $X_h = \sum_{i \in U_h} x_i$ of a variable x for each area h. We assume that there is a survey that measures x_i from the sample but its sample size is not large enough to obtain estimates for X_h with reasonable accuracy. Consider one of the surveys, called survey A, as the main survey, and let $\hat{X}_h$ denote a design-consistent estimator of X_h obtained

1. Jae-kwang Kim, Department of Statistics, Iowa State University, Ames, Iowa, 50011, U.S.A.; Seunghwan Park, Department of Statistics, Seoul National University, Seoul, 151-747, Korea. E-mail: kkampsh@gmail.com; Seo-young Kim, Statistical Research Institute, Statistics Korea, Daejon, 302-847, Korea.

from survey A. Often, we compute $\hat{X}_h = \sum_{i \in A_h} w_{ia} x_i$, where A_h is the set of sample A for subpopulation h and w_{ia} is the weight of unit i in sample A.

In addition to the main survey, suppose that there is another survey, called survey B, that measures a rough estimate for x_i. Let y_{1i} be the measurement taken from survey B. We may assume that y_{1i} is a rough measurement of x_i with some level of measurement error. Thus, we may assume

$$y_{1i} = \beta_0 + \beta_1 x_i + e_{1i} \tag{1.1}$$

for some (β_0, β_1), where $e_{1i} \sim (0, \sigma_{e1}^2)$. Model (1.1) is variable-specific and the linear regression assumption or equal variance assumptions can be relaxed later. If $(\beta_0, \beta_1) = (0,1)$, then model (1.1) means that there is no measurement bias. Note that model parameters (β_0, β_1) in (1.1) are not area specific, but may be different for groups of areas, as demonstrated in the Korean labor force survey application in Section 5. Separate regression models for different groups may lead to smaller model errors and thus improve the statistical efficiency of the proposed method. From survey B, we can obtain another estimator $\hat{Y}_{1h} = \sum_{i \in B_h} w_{ib} y_{1i}$ of X_h, where w_{ib} is the weight of unit i in the sample from survey B and B_h is the B-sample for subpopulation h. Note that $\hat{Y}_{1h}$ can be obtained, for each area, if the same areas are identified in both surveys A and B. Model (1.1) can be used to combine information from the two surveys.

Finally, another source of information can be the Census information. Census information does not suffer from coverage error or sampling error. But, it may have measurement errors and it does not provide updated information for each month or year. Let y_{2i} be the measurement for unit i from the Census. The subpopulation total $Y_{2h} = \sum_{i \in C_h} y_{2i}$ is available when C_h is the set of Census C for subpopulation h.

Table 1.1 summarizes the major sources of information that we can consider into small area estimation.

Table 1.1
Available information for small area estimation

Data	Observation	Area level estimate	Properties
Survey A	direct obs. (x_i)	$\hat{X}_h, \hat{V}(\hat{X}_h)$	Sampling error (large)
Survey B	aux. obs. (y_{1i})	$\hat{Y}_{1h}, \hat{V}(\hat{Y}_{1h})$	Bias Measurement error Sampling error
Census	aux. obs. (y_{2i})	Y_{2h}	Measurement error No updated information

In this paper, we consider an area-level model approach for small area estimation combining all available information. The proposed approach is based on the measurement error models, where the sampling errors of the direct estimators are treated as measurement errors, and all the other auxiliary information are combined through a set of linking models. The proposed approach is applied to the small area estimation problem for labor force surveys in Korea, where three estimates are combined to produce small area estimates for unemployment rates.

The paper is organized as follows. In Section 2, the basic setup is introduced and the small area estimation problem is viewed as a measurement error model prediction problem. In Section 3, parameter estimation for the area level small area model is discussed. In Section 4, estimation of mean squared error is briefly discussed. In Section 5, the proposed method is applied to the labor force survey data in Korea. Concluding remarks are made in Section 6.

2 Basic theory

In this section, we first introduce the basic theory for combining the information for small area estimation. We first consider the simple case of combining two surveys. Assume that there are two surveys, survey A and survey B, obtained from separate probability sampling designs. The two surveys are not necessarily independent. From survey A, we obtain a design unbiased estimator $\hat{X}_{h,a} = \sum_{i \in A_h} w_{ia} x_i$ and its variance estimator $\hat{V}\left(\hat{X}_h\right)$. From survey B, we obtain a design unbiased estimator $\hat{Y}_{1h} = \sum_{i \in B_h} w_{ib} y_{1i}$ of $Y_{1h} = \sum_{i \in U_h} y_{1i}$. The sampling error of $\left(\hat{X}_h, \hat{Y}_{1h}\right)$ can be expressed by the *sampling error model*

$$\begin{pmatrix} \hat{X}_h \\ \hat{Y}_{1h} \end{pmatrix} = \begin{pmatrix} X_h \\ Y_{1h} \end{pmatrix} + \begin{pmatrix} N_h a_h \\ N_h b_h \end{pmatrix} \tag{2.1}$$

and a_h and b_h represent the sampling errors associated with $\hat{X}_h / N_h$ and $\hat{Y}_{1h} / N_h$ such that

$$\begin{pmatrix} a_h \\ b_h \end{pmatrix} \sim \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} V(a_h) & \text{Cov}(a_h, b_h) \\ \text{Cov}(a_h, b_h) & V(b_h) \end{pmatrix} \right].$$

Our parameter of interest is the population total X_h of x in area h.

From (1.1), we obtain the following area level model:

$$Y_{1h} = N_h \beta_0 + \beta_1 X_h + \tilde{e}_{1h}, \tag{2.2}$$

where $(N_h, X_h, Y_{1h}, \tilde{e}_{1h}) = \sum_{i \in U_h} (1, x_i, y_{1i}, e_{1i})$. We can express (2.2) in terms of population mean

$$\bar{Y}_{1h} = \beta_0 + \bar{X}_h \beta_1 + \bar{e}_{1h}, \tag{2.3}$$

where $(\bar{X}_h, \bar{Y}_{1h}, \bar{e}_{1h}) = N_h^{-1} \sum_{i \in U_h} (x_i, y_{1i}, e_{1i})$. If we use a nested error model

$$e_{1hi} = \varepsilon_h + u_{hi} \tag{2.4}$$

where $\varepsilon_h \sim \left(0, \sigma_e^2\right)$ and $u_{hi} \sim \left(0, \sigma_u^2\right)$, then $\bar{e}_{1h} \sim \left(0, \sigma_{e,h}^2\right)$, $\sigma_{e,h}^2 = \sigma_e^2 + \sigma_u^2 / N_h$. The nested error model is quite popular in small area estimation (e.g., Battese, Harter and Fuller 1988) and it assumes that $\text{Cov}\left(e_{1hi}, e_{1hj}\right) = \sigma_e^2$ for $i \neq j$. Because N_h is often quite large, we can safely assume that $\bar{e}_{1h} \sim \left(0, \sigma_{e,h}^2 = \sigma_e^2\right)$. The model (2.2) is called *structural error model* because it describes the structural

relationship between the two latent variables Y_{1h} and X_h. The two models, (2.1) and (2.2), are often encountered in the measurement error model literature (Fuller 1987). Thus, the model for small area estimation can be viewed as a measurement error model, as suggested by Fuller (1991) who originally used the measurement error model approach in the unit-level modeling for small area estimation.

Now, if we define $(\bar{y}_{1h}, \bar{x}_h) = N_h^{-1}\left(\hat{Y}_{1h}, \hat{X}_h\right)$, combining (2.1) and (2.3), we have

$$\begin{pmatrix} \bar{y}_{1h} \\ \bar{x}_h \end{pmatrix} = \begin{pmatrix} \beta_0 & \beta_1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ \bar{X}_h \end{pmatrix} + \begin{pmatrix} b_h + \bar{e}_{1h} \\ a_h \end{pmatrix}$$

which can also be written as

$$\begin{pmatrix} \bar{y}_{1h} - \beta_0 \\ \bar{x}_h \end{pmatrix} = \begin{pmatrix} \beta_1 \\ 1 \end{pmatrix} \bar{X}_h + \begin{pmatrix} b_h + \bar{e}_{1h} \\ a_h \end{pmatrix}. \tag{2.5}$$

Thus, when all the model parameters in (2.5) are known, the best estimator of $\bar{X}_h$ can be computed by

$$\hat{\bar{X}}_h = \left\{(\beta_1, 1) V_h^{-1} (\beta_1, 1)'\right\}^{-1} (\beta_1, 1) V_h^{-1} (\bar{y}_{1h} - \beta_0, \bar{x}_h)' \tag{2.6}$$

where V_h is the variance-covariance matrix of $(b_h + \bar{e}_{1h}, a_h)'$. The variance of $\hat{\bar{X}}_h$ is given by $\left\{(\beta_1, 1) V_h^{-1} (\beta_1, 1)'\right\}^{-1}$. The estimator in (2.6) can be called the Generalized Least Squares (GLS) estimator because it uses the technique of the generalized least squares method in the linear model theory. The GLS method is useful because it is optimal and it can incorporate additional sources of information naturally. For example, if another estimator $\bar{y}_{2h}$ for $\bar{Y}_{2h}$ is also available and satisfies

$$\bar{Y}_{2h} = \gamma_0 + \gamma_1 \bar{X}_h + \bar{e}_{2h}$$

and

$$\bar{y}_{2h} = \bar{Y}_{2h} + c_h,$$

then the extended GLS model is written as

$$\begin{pmatrix} \bar{y}_{2h} - \gamma_0 \\ \bar{y}_{1h} - \beta_0 \\ \bar{x}_h \end{pmatrix} = \begin{pmatrix} \gamma_1 \\ \beta_1 \\ 1 \end{pmatrix} \bar{X}_h + \begin{pmatrix} c_h + \bar{e}_{2h} \\ b_h + \bar{e}_{1h} \\ a_h \end{pmatrix} \tag{2.7}$$

and the GLS estimator can be obtained by

$$\hat{\bar{X}}_{h2} = \left\{(\gamma_1, \beta_1, 1) V_{h2}^{-1} (\gamma_1, \beta_1, 1)'\right\}^{-1} (\gamma_1, \beta_1, 1) V_{h2}^{-1} (\bar{y}_{2h} - \gamma_0, \bar{y}_{1h} - \beta_0, \bar{x}_h)'$$

where V_{h2} is the variance-covariance matrix of $(c_h + \bar{e}_{2h}, b_h + \bar{e}_{1h}, a_h)'$. The GLS estimator has variance $\left\{(\gamma_1, \beta_1, 1) V_{h2}^{-1} (\gamma_1, \beta_1, 1)'\right\}^{-1}$. If $\bar{y}_{2h}$ is independent of $(\bar{x}_h, \bar{y}_{1h})$, the efficiency gain by incorporating $\bar{y}_{2h}$ into GLS in terms of relative variance can be expressed as

$$\frac{V\left(\hat{\bar{X}}_{h2}\right) - V\left(\hat{\bar{X}}_{h}\right)}{V\left(\hat{\bar{X}}_{h}\right)} = -\frac{\{V(\bar{y}_{2h}/\gamma_1)\}^{-1}}{\left\{V\left(\hat{\bar{X}}_{h}\right)\right\}^{-1} + \{V(\bar{y}_{2h}/\gamma_1)\}^{-1}},$$

where $V(\bar{y}_{2h}/\gamma_1) = V(c_h + \bar{e}_{2h})/\gamma_1^2$. The gain is high if both the sampling variance of $\bar{y}_{2h}$ and the model variance $V(\bar{e}_{2h})$ are small. If $\gamma_1 = 0$, then there is no gain.

Remark 1 *Note that model (2.5) can also be written as*

$$\begin{pmatrix} \beta_1^{-1}(\bar{y}_{1h} - \beta_0) \\ \bar{x}_h \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \bar{X}_h + \begin{pmatrix} (b_h + \bar{e}_{1h})/\beta_1 \\ a_h \end{pmatrix}. \tag{2.8}$$

The GLS estimator obtained from (2.8), which is the same as the GLS estimator obtained from (2.5), can be expressed as

$$\hat{\bar{X}}_h = \alpha_h \bar{x}_h + (1 - \alpha_h)\tilde{x}_h \tag{2.9}$$

where $\tilde{x}_h = \beta_1^{-1}(\bar{y}_{1h} - \beta_0)$ *and*

$$\begin{aligned} \alpha_h &= \frac{V(\tilde{x}_h) - \mathrm{Cov}(\bar{x}_h, \tilde{x}_h)}{V(\bar{x}_h) + V(\tilde{x}_h) - 2\mathrm{Cov}(\bar{x}_h, \tilde{x}_h)} \\ &= \frac{\sigma_{e,h}^2 + V(b_h) - \beta_1 \mathrm{Cov}(a_h, b_h)}{\sigma_{e,h}^2 + V(b_h) + \beta_1^2 V(a_h) - 2\beta_1 \mathrm{Cov}(a_h, b_h)}. \end{aligned}$$

The estimator $\tilde{x}_h$, *when computed with estimated parameter* $\hat{\beta} = \left(\hat{\beta}_0, \hat{\beta}_1\right)$, *is called the synthetic estimator and the optimal estimator in (2.9) is often called the composite estimator. It can be shown that, ignoring the effect of estimating* β, *the variance of the composite estimator is equal to*

$$V\left(\hat{\bar{X}}_h - \bar{X}_h\right) = \alpha_h V(\bar{x}_h) + (1 - \alpha_h)\mathrm{Cov}(\bar{x}_h, \tilde{x}_h) \tag{2.10}$$

and, as $\alpha_h < 1$, *the composite estimator is more efficient than the direct estimator.*

3 Parameter estimation

Now, we discuss estimation of the model parameters in (2.3). The GLS estimator of $\beta = (\beta_0, \beta_1)$ can be obtained by minimizing

$$Q^*(\beta_0, \beta_1) = \sum_{h=1}^{H} \frac{(\bar{y}_{1h} - \beta_0 - \beta_1 \bar{x}_h)^2}{V(\bar{y}_{1h} - \beta_0 - \beta_1 \bar{x}_h)}. \tag{3.1}$$

Since

$$V(\bar{y}_{1h} - \beta_0 - \bar{x}_h\beta_1) = \sigma_{e,h}^2 + (-\beta_1, 1)\Sigma_h(-\beta_1, 1)', \tag{3.2}$$

where $\sigma_{e,h}^2 = V(\bar{e}_{1h})$ and $\Sigma_h = V\left\{(a_h, b_h)'\right\}$, we can express

$$Q^*(\beta_0, \beta_1) = \sum_{h=1}^{H} w_h(\beta_1)(\bar{y}_{1h} - \beta_0 - \beta_1 \bar{x}_h)^2, \tag{3.3}$$

where $w_h(\beta_1) = \left\{\sigma_{e,h}^2 + (-\beta_1, 1)\Sigma_h(-\beta_1, 1)'\right\}^{-1}$. Now, by solving $\partial Q^* / \partial \beta = 0$, we have

$$\hat{\beta}_0 = \bar{y}_w - \hat{\beta}_1 \bar{x}_w \tag{3.4}$$

and

$$\hat{\beta}_1 = \frac{\sum_{h=1}^{H} w_h\left(\hat{\beta}_1\right)\{(\bar{x}_h - \bar{x}_w)(\bar{y}_{1h} - \bar{y}_{1w}) - C(a_h, b_h)\}}{\sum_{h=1}^{H} w_h\left(\hat{\beta}_1\right)\left\{(\bar{x}_h - \bar{x}_w)^2 - V(a_h)\right\}}, \tag{3.5}$$

where

$$(\bar{x}_w, \bar{y}_w) = \left\{\sum_{h=1}^{H} w_h\left(\hat{\beta}_1\right)\right\}^{-1} \sum_{h=1}^{H} w_h\left(\hat{\beta}_1\right)(\bar{x}_h, \bar{y}_h).$$

Note that the weight $w_h(\beta_1)$ depends on β_1. Thus, the solution (3.5) can be obtained by an iterative algorithm. Once $\hat{\beta}_1$ is computed by (3.5), then $\hat{\beta}_0$ is obtained by (3.4).

Now, we discuss the estimation of model variance $\sigma_{e,h}^2$. The simplest method is the Method of Moments (MOM). That is, we can use

$$E\left\{(\bar{y}_{1h} - \beta_0 - \bar{x}_h\beta_1)^2 - \beta_1^2 V(a_h) + 2\beta_1 C(a_h, b_h) - V(b_h)\right\} = \sigma_{e,h}^2 \tag{3.6}$$

to obtain an unbiased estimator of $\sigma_{e,h}^2$. Under the nested error model in (2.4), we have $\sigma_{e,h}^2 = \sigma_e^2$ and

$$E\left\{(\bar{y}_{1h} - \beta_0 - \bar{x}_h\beta_1)^2 - \beta_1^2 V(a_h) + 2\beta_1 C(a_h, b_h) - V(b_h)\right\} = \sigma_e^2. \tag{3.7}$$

Thus, similarly to Fuller (2009), the MOM estimator of σ_e^2 can be obtained by

$$\hat{\sigma}_e^2 = \sum_{h=1}^{H} \kappa_h \left\{\left(\bar{y}_{1h} - \hat{\beta}_0 - \bar{x}_h\hat{\beta}_1\right)^2 - \left(-\hat{\beta}_1, 1\right)\Sigma_h\left(-\hat{\beta}_1, 1\right)\right\} \tag{3.8}$$

where

$$\kappa_h \propto \left\{\hat{\sigma}_e^2 + \left(-\hat{\beta}_1, 1\right)\Sigma_h\left(-\hat{\beta}_1, 1\right)\right\}^{-1}$$

and $\sum_{h=1}^{H} \kappa_h = 1$. Because κ_h depends on $\hat{\sigma}_e^2$, the solution (3.8) can be obtained iteratively, using $\hat{\sigma}_e^2 = 0$ as an initial value. Fay and Herriot (1979) used an alternative method which is based on the iterative solution to nonlinear equation:

$$\sum_{h=1}^{H} \frac{\left(\bar{y}_{1h} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x}_h\right)^2}{\sigma_e^2 + \left(-\hat{\beta}_1, 1\right) \Sigma_h \left(-\hat{\beta}_1, 1\right)'} = H - 2.$$

Writing the above equation as $g\left(\sigma_e^2\right) = H - 2$, a Newton-type method for $g(\theta) = 0$ with $\theta = \sigma_e^2$ can be obtained by

$$\theta^{(t+1)} = \theta^{(t)} + \frac{1}{g'\left(\theta^{(t)}\right)}\left(H - 2 - g\left(\theta^{(t)}\right)\right) \tag{3.9}$$

where

$$g'(\theta) = -\sum_{h=1}^{H} \frac{\left(\bar{y}_{1h} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x}_h\right)^2}{\left\{\theta + \left(-\hat{\beta}_1, 1\right) \Sigma_h \left(-\hat{\beta}_1, 1\right)'\right\}^2}.$$

Assuming $\sigma_{e,h}^2 \equiv \sigma_e^2$, we now describe the whole parameter estimation procedure as follows:

Step 1 Compute the initial estimator of (β_0, β_1) by setting $\hat{\sigma}_e^2 = 0$ in (3.4) and (3.5).

Step 2 Based on the current value of $\left(\hat{\beta}_0, \hat{\beta}_1\right)$, compute $\hat{\sigma}_e^2$ using the iterative algorithm in (3.9).

Step 3 Use the current value of $\hat{\sigma}_e^2$, compute the updated estimator of (β_0, β_1) by (3.4) and (3.5).

Step 4 Repeat [Step 2]-[Step 3] until convergence.

The proposed parameter estimation method estimates $\beta = (\beta_0, \beta_1)$ by the GLS and estimates σ_e^2 by the MOM iteratively. Note that the estimation of β is based on data from all areas. If separate regression models are used, then the proposed parameter estimation method can be applied to the groups of areas. Instead of this separate iterative estimation method, we can also consider another method based on maximum likelihood estimation (MLE) under parametric distributional assumptions. See Carroll, Rupert, and Stefanski (1995) and Schafer (2001) for further discussion of MLE for parameters in the measurement error models.

Remark 2 *If $\sigma_{e,h}^2 = \sigma_e^2$ is not true, we can consider some alternative model such as*

$$\bar{e}_h \sim \left(0, \bar{X}_h \sigma_e^2\right). \tag{3.10}$$

To check whether model (3.10) holds, one can compute

$$\nu_h = \left(\bar{y}_{1h} - \hat{\beta}_0 - \bar{x}_h \hat{\beta}_1\right)^2 - \hat{\beta}_1^2 V(a_h) + 2\hat{\beta}_1 \hat{C}(a_h, b_h) - V(b_h) \tag{3.11}$$

and plot ν_h on $\bar{x}_h$. If the plot shows a linear relationship, then (3.10) can be treated as a reasonable model. Under model (3.10), we can obtain σ_e^2 by a ratio method:

$$\hat{\sigma}_e^2 = \frac{\sum_{h=1}^{H} \kappa_h \nu_h}{\sum_{h=1}^{H} \kappa_h \hat{\bar{X}}_h} \tag{3.12}$$

where

$$\kappa_h \propto \left\{ \hat{\bar{X}}_h \hat{\sigma}_e^2 + \left(-\hat{\beta}_1, 1\right) \Sigma_h \left(-\hat{\beta}_1, 1\right) \right\}^{-1}$$

with $\sum_{h=1}^{H} \kappa_h = 1$, $\hat{\bar{X}}_h$ *is defined in (2.9), and* ν_h *is defined in (3.11). Because* κ_h *also depends on* σ_e^2, *the solution (3.12) can be obtained iteratively.*

Remark 3 *We can also consider a transformation* $\bar{x}_h^* = T(\bar{x}_h)$ *and* $\bar{y}_{1h}^* = T(\bar{y}_{1h})$ *to improve the approximation to asymptotic normality. To check the departure from normality, plot* $n_{ha} \bar{V}(\bar{x}_h)$ *on* $\bar{x}_h$. *If the plot shows some structural relationship of* $\bar{x}_h$ *then the normality assumption can be doubted. Now, consider the following transformation*

$$T(x) = \log(x). \tag{3.13}$$

Note that the asymptotic variance of $\bar{x}_h^* = T(\bar{x}_h)$ *is equal to*

$$V\left(\bar{x}_h^*\right) \doteq \frac{1}{\left(\bar{x}_h\right)^2} V(\bar{x}_h).$$

Such transformation is a variance stabilizing transformation and is useful when we want to improve the approximation to normality.

Once the GLS estimator $\hat{\bar{X}}_h^*$ *of* $\bar{X}_h^*$ *is obtained, then we need to apply the inverse transformation to obtain the best estimator of* $\bar{X}_h = T^{-1}\left(\bar{X}_h^*\right) := Q\left(\bar{X}_h^*\right)$. *Simply applying the inverse transformation will lead to biased estimation. To correct for the bias, we can use a second-order Taylor linearization. Using a Taylor expansion, we have*

$$Q\left(\hat{\bar{X}}_h^*\right) \doteq Q\left(\bar{X}_h^*\right) + Q'\left(\bar{X}_h^*\right)\left(\hat{\bar{X}}_h^* - \bar{X}_h\right) + \frac{1}{2} Q''\left(\bar{X}_h^*\right)\left(\hat{\bar{X}}_h^* - \bar{X}_h\right)^2$$

and so, if we use $Q\left(\hat{\bar{X}}_h^*\right)$ *as an estimator for* $\bar{X}_h = Q\left(\bar{X}_h^*\right)$, *we have, ignoring the smaller order terms,*

$$E\left\{Q\left(\hat{\bar{X}}_h^*\right)\right\} = \bar{X}_h + \frac{1}{2} Q''\left(\bar{X}_h^*\right) V\left(\hat{\bar{X}}_h^*\right).$$

For the transformation in (3.13), we have $Q\left(\bar{X}_h^*\right) = \exp\left(\bar{X}_h^*\right)$ *and so* $Q''\left(\bar{X}_h^*\right) = \bar{X}_h$. *Thus,* $\hat{\bar{X}}_h = Q\left(\hat{\bar{X}}_h^*\right)$, *we have*

$$E\left(\hat{\bar{X}}_h\right) \cong \bar{X}_h + \frac{1}{2} \bar{X}_h V\left(\hat{\bar{X}}_h^*\right)$$

and the bias-corrected estimator of $\bar{X}_h$ *is*

$$\hat{\bar{X}}_{h,bc} = \frac{\hat{\bar{X}}_h}{1 + 0.5V\left(\hat{\bar{X}}_h^*\right)}, \tag{3.14}$$

where $V\left(\hat{\bar{X}}_h^*\right)$ *is computed by the MSE estimation method which will be discussed in Section 4.*

4 MSE estimation

We now discuss mean squared error (MSE) estimation of the GLS estimator $\hat{\bar{X}}_h$ which is given by (2.9). Note that the GLS estimator is a function of (β_0, β_1) and σ_e^2. If the model parameters are known, then the MSE of $\hat{\bar{X}}_h$ is equal to $M_{h1} = \alpha_h V(\bar{x}_h) + (1 - \alpha_h)\mathrm{Cov}(\bar{x}_h, \tilde{x}_h)$, as discussed in Remark 1. That is, writing $\theta = \left(\beta_0, \beta_1, \sigma_e^2\right)$ and $\hat{\bar{X}}_h = \hat{\bar{X}}_h(\theta)$, the actual prediction for $\bar{X}_h$ is computed by $\hat{\bar{X}}_{eh} = \hat{\bar{X}}_h\left(\hat{\theta}\right)$. To account for the effect of estimating the model parameters, we first note the following decomposition of $\mathrm{MSE}\left(\hat{\bar{X}}_h^*\right)$:

$$\begin{aligned} \mathrm{MSE}\left(\hat{\bar{X}}_{eh}\right) &= \mathrm{MSE}\left(\hat{\bar{X}}_h\right) + E\left\{\left(\hat{\bar{X}}_{eh} - \hat{\bar{X}}_h\right)^2\right\} \\ &=: M_{h1} + M_{h2}, \end{aligned}$$

which was originally proved by Kackar and Harville (1984) under normality assumptions. The first term, M_{h1}, is of order $1/n_h$, where n_h is the size of A_h, and the second term, M_{h2}, is of order $1/n$ with $n = \sum_{h=1}^{H} n_h$. The second term is often much smaller than the first term.

We consider a jackknife approach to estimate the MSE. Use of the jackknife for bias-corrected estimation was originally proposed by Quenouille (1956). Jiang, Lahiri and Wan (2002) provided a rigorous justification of the jackknife method for the MSE estimation in small area estimation. The following steps can be used for the jackknife computation.

Step 1 Calculate the k^{th} replicate $\hat{\theta}^{(-k)}$ of $\hat{\theta}$ by deleting the k^{th} area data set $(\bar{x}_k, \bar{y}_{1k})$ from the full data set $\{(\bar{x}_h, \bar{y}_{1h}); h = 1, 2, \ldots, H\}$. This calculation is done for each k to get H replicates of θ: $\left\{\hat{\theta}^{(-k)}; k = 1, \ldots, H\right\}$ which, in turn, provide H replicates of $\hat{\bar{X}}_h$: $\left\{\hat{\bar{X}}_h^{(-k)}; k = 1, 2, \ldots, H\right\}$, where $\hat{\bar{X}}_h^{(-k)} = \hat{\bar{X}}_h\left(\hat{\theta}^{(-k)}\right)$.

Step 2 Calculate the estimator of M_{h2} as

$$\hat{M}_{2h} = \frac{H-1}{H}\sum_{k=1}^{H}\left(\hat{\bar{X}}_h^{(-k)} - \hat{\bar{X}}_h\right)^2. \tag{4.1}$$

Step 3 Calculate the estimator of M_{h1} as

$$\hat{M}_{1h} = \hat{\alpha}_h^{(\mathrm{JK})} V(\bar{x}_h) + \left(1 - \hat{\alpha}_h^{(\mathrm{JK})}\right)\mathrm{Cov}(\bar{x}_h, \tilde{x}_h) \tag{4.2}$$

where $\hat{\alpha}_h^{(\mathrm{JK})}$ is a bias-corrected estimator of α_h given by

$$\begin{aligned}\hat{\alpha}_h^{(\mathrm{JK})} &= \hat{\alpha}_h - \frac{H-1}{H}\sum_{k=1}^{H}\left(\hat{\alpha}_h^{(-k)} - \hat{\alpha}_h\right),\\ \hat{\alpha}_h &= \frac{\hat{\sigma}_e^2 + V(b_h) - \hat{\beta}_1 \mathrm{Cov}(a_h, b_h)}{\hat{\sigma}_e^2 + V(b_h) + \hat{\beta}_1^2 V(a_h) - 2\hat{\beta}_1 \mathrm{Cov}(a_h, b_h)},\end{aligned}$$

and

$$\hat{\alpha}_h^{(-k)} = \frac{\hat{\sigma}_e^{(-k)2} + V(b_h) - \hat{\beta}_1^{(-k)} \mathrm{Cov}(a_h, b_h)}{\hat{\sigma}_e^{(-k)2} + V(b_h) + \left(\hat{\beta}_1^{(-k)}\right)^2 V(a_h) - 2\hat{\beta}_1^{(-k)} \mathrm{Cov}(a_h, b_h)}.$$

Remark 4 *For the transformation in (3.13), we use the bias-corrected estimator in (3.14) and its MSE estimation method needs to be changed. Using* $\hat{\bar{X}}_{eh,bc}$ *to denote the bias-corrected estimator in (3.14) evaluated at* $\hat{\theta}$*, we can have the*

$$\begin{aligned}\mathrm{MSE}\left(\hat{\bar{X}}_{eh,bc}\right) &= \mathrm{MSE}\left(\hat{\bar{X}}_{eh}\right)\\ &= \mathrm{MSE}\left\{Q\left(\hat{\bar{X}}_{eh}^*\right)\right\}\\ &\cong \left\{Q'\left(\bar{X}_h^*\right)\right\}^2 \cdot \mathrm{MSE}\left(\hat{\bar{X}}_{eh}^*\right)\\ &= \bar{X}_h^2 \cdot \mathrm{MSE}\left(\hat{\bar{X}}_{eh}^*\right),\end{aligned}$$

where the first equality follows that $\hat{\bar{X}}_{h,bc} - \hat{\bar{X}}_h$ *is of order* $O_p\left(n_h^{-1}\right)$*. The MSE of* $\hat{\bar{X}}_h^*$*, the EGLS estimator of* $\bar{X}_h^*$ *after transformation, is computed by (4.1) and (4.2). Once* $\mathrm{MSE}\left(\hat{\bar{X}}_{eh}^*\right)$ *is estimated, we should multiply it by* $\hat{\bar{X}}_h^2$ *to obtain the MSE estimator of the back-transformed EGLS estimator* $\hat{\bar{X}}_{eh,bc}$*.*

5 Application to Korean Labor Force survey

We now consider an application of the proposed method to the labor force surveys in Korea. In Korea, two different labor force surveys are used to obtain information about employment. One is the Korean Labor Force (KLF) survey and the other is the Local Area labor force (LALF) survey. The KLF survey has about 7K sample households but LALF has about 200K sample households. Because LALF is a large-scale survey employing a lot of part time interviewers, there is a certain level of measurement errors in the LALF survey. We assume that the KLF has no measurement error, although it has significant sampling errors at the small area level. The KLF sample is a second-phase sample from the LALF sample. Thus, the sampling errors for two survey estimates are correlated. Let $\bar{X}_h$ be the (true) unemployment rate for area h. The small area level we considered is called “Gu”. The number of “Gu” in Korea is 229.

We observe $\bar{x}_h$ from KLF survey and $\bar{y}_{1h}$ from the LALF survey. To construct linking models, we first partition the population into two regions, urban region and rural region, based on the proportion of the households working on agricultural practice. Within each region, we build models separately (same model but allows for different parameter) and estimate the model parameters separately. The structural model is

$$\bar{Y}_h = \beta_1 \bar{X}_h + e_h \tag{5.1}$$

with $e_h \sim (0, \sigma_e^2)$. Here, we set $\beta_0 = 0$ to guarantee that the GLS estimator of $\bar{X}_h$ is nonnegative. The sampling error model remains the same. In this case, β_1 can be estimated by

$$\hat{\beta}_1 = \frac{\sum_{h=1}^{H} w_h(\hat{\beta}_1)\{\bar{x}_h \bar{y}_{1h} - C(a_h, b_h)\}}{\sum_{h=1}^{H} w_h(\hat{\beta}_1)\{\bar{x}_h^2 - V(a_h)\}}. \tag{5.2}$$

The sampling variance of (a_h, b_h) is computed using the method of reversed two-phase sampling described in the Appendix. The model variance is estimated by the method of moment technique in (3.8) with $\hat{\beta}_0 = 0$. The GLS estimator can be computed by (2.9) with $\tilde{x}_h = \hat{\beta}_1^{-1} \bar{y}_{1h}$.

In addition to the two surveys, we can also use the Census information. The GLS model incorporating the three sources of information can be expressed as

$$\begin{pmatrix} \bar{Y}_{2h} \\ \bar{y}_{1h} \\ \bar{x}_h \end{pmatrix} = \begin{pmatrix} \gamma_1 \\ \beta_1 \\ 1 \end{pmatrix} \bar{X}_h + \begin{pmatrix} \bar{e}_{2h} \\ b_h + \bar{e}_{1h} \\ a_h \end{pmatrix}$$

where $\bar{Y}_{2h}$ is the census result for area h. Because the Census estimate does not suffer from sampling error, we have only model error e_{2h} which represents the error when we model $E(\bar{Y}_{h2}) = \gamma_1 \bar{X}_h$. The model parameters can be obtained using the method in Section 3 with $\Sigma_h = \text{diag}(0, V(a_h, b_h))$. The GLS estimator of $\bar{X}_h$ can be obtained easily. The MSE part can be computed by using the fact that

$$V\left(\hat{\bar{X}}_h - \bar{X}_h\right) = \begin{pmatrix} \gamma_1 \\ \beta_1 \\ 1 \end{pmatrix}' \left\{ V \begin{pmatrix} \bar{e}_{2h} \\ b_h + \bar{e}_{1h} \\ a_h \end{pmatrix} \right\}^{-1} \begin{pmatrix} \gamma_1 \\ \beta_1 \\ 1 \end{pmatrix} := M_{h1}$$

and applying the jackknife method for bias correction.

Figure 5.1 presents the plot of the unemployment rate of KLF against LALF for urban areas. From Figure 5.1, we can find that there is a linear structural relationship between KLF and LALF. Instead of the usual residual $\hat{\bar{e}}_h$ in the structural error model, $\hat{v}_h$ are used as the residuals in the regression model with measurement errors, where $\hat{v}_h = \bar{y}_{1h} - \hat{\beta}_1 \bar{x}_h$. Figure 5.2 contains a plot of $\hat{v}_h$ against $\hat{\bar{X}}_h$ for urban area. The plot shows that the assumption of equal variance σ_e^2 is slightly violated. The heteroscedastic variance model in Remark 2 was also considered but the results did not change significantly.

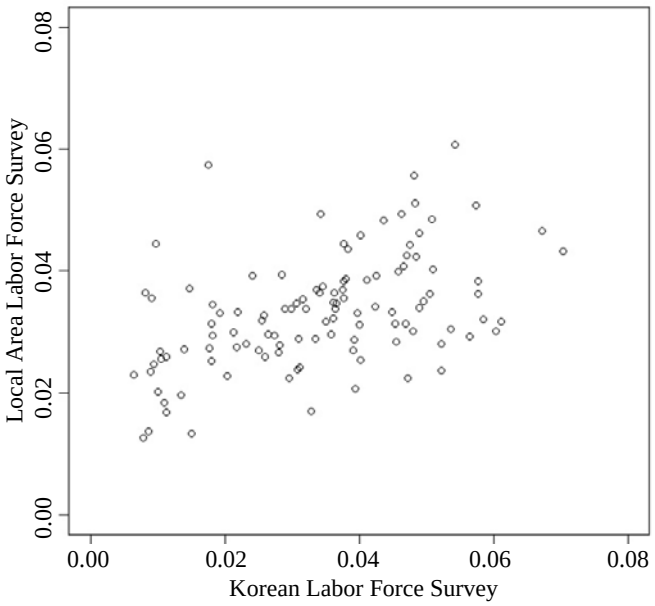


Figure 5.1 Plot of unemployment rate for KLF and LALF survey for urban area.

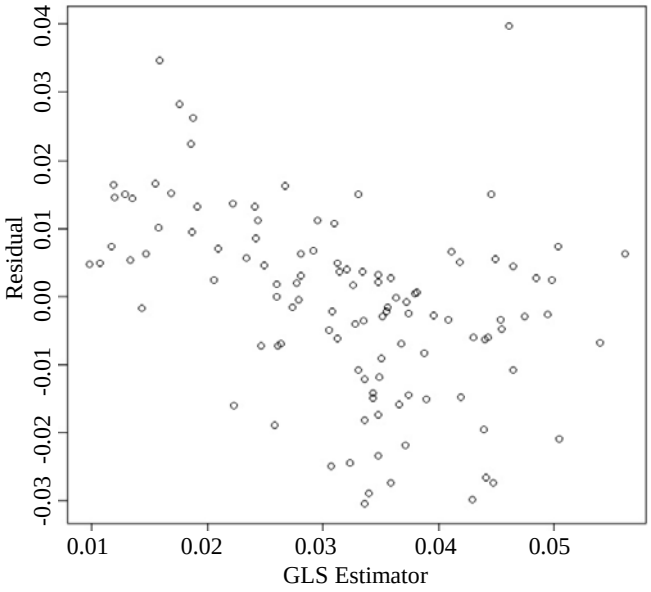


Figure 5.2 Plot of residuals against estimated values for urban area.

Table 5.1 presents the performance of the small area estimates in terms of the MSE estimates. We considered four different estimators of $\bar{X}_h$. KLF represents the result derived using only Korea Labor Force survey, LALF represents the result using only Local Area Labor Force survey, GLS 1 represents the

result for combining both surveys KLF and LALF, and GLS 2 represents the result for combining KLF, LALF and the Census data. Table 5.1 shows that the GLS 2 method provides the smallest mean squared errors.

Table 5.1
Quartile of the MSE performance of the small area estimates for the 229 areas

MSE	1st Q	Median	3rd Q	Mean
KLF	0.0000630	0.0001210	0.0002395	0.0002476
LALF	0.0001123	0.0001330	0.0001695	0.0001482
GLS 1	0.0000444	0.0000738	0.0001210	0.0000893
GLS 2	0.0000405	0.0000543	0.0000721	0.0000575

6 Concluding remark

In this paper, a small area estimation problem is treated as a measurement error model prediction problem where the covariates, which are the direct estimates for small areas, are subject to sampling errors. In our measurement error model approach, the sampling errors of the direct estimators are treated as measurement errors and the structural error model can be used to link the other auxiliary estimates to the direct estimators. The proposed model is actually the opposite of the model of Ybarra and Lohr (2008), where the direct estimator is treated as a dependent variable in the regression model and the nonsampling errors of auxiliary estimates are treated as measurement errors.

In our approach, each auxiliary estimate is treated as a dependent variable in the regression model using the direct estimate as the covariate and the sampling error of the direct estimator is treated as measurement error. The measurement error variance is easy to estimate because it is essentially the sampling variance of the direct estimate. The measurement error model approach is also very useful when there are several sources of auxiliary information of area-levels. Unlike the Bayesian approach, the resulting estimator does not rely on parametric model assumptions about the structural error model and is still optimal in the sense of minimizing the mean squared errors among the class of unbiased estimators that are linear in the available data.

In the example of the Korean labor survey application, two sample estimates and the Census information are used to compute the GLS estimates for small area parameters and the two sample estimates are correlated due to the two-phase sampling structure. We simply used linear regression models for the linking models, mainly for the sake of computational simplicity. Instead of the linear model, one may consider a generalized linear model to improve model prediction power. Such extension would involve the theory for nonlinear measurement error models. Further investigation on this extension will be a topic of future research.

Acknowledgements

We thank an anonymous referee and the Associate Editor for their constructive comments. The research of the first author was partially supported by a grant from NSF (MMS-121339).

Appendix

Reversed two-phase sampling

In the classical two-phase sampling, the second-phase sample (A_2) is a subset of the first-phase sample (A_1). We consider another type of sampling design that has a reversed structure of the two-phase sampling design. In the reversed two-phase sampling design, we have the following sampling steps:

Step 1 From the finite population, we select the first-phase sample A_1 of size n_1.

Step 2 In the second-phase sample, we select A_2 from $U - A_1$ of size n_2. The final sample A consists of A_1 and A_2. That is, $A = A_1 \cup A_2$ and $|A| = n = n_1 + n_2$.

The reversed two-phase sampling is used when the sample is augmented by an additional sampling procedure.

To discuss parameter estimation under reversed two-phase sampling, let $\pi_{1i} = \Pr(i \in A_1)$ be the first-order inclusion probability for A_1. Let $\pi_{2i|1} = \Pr\left(i \in A_2 | A_1^c\right)$ be the conditional first-order inclusion probability for A_2 given $A_1^c = U - A_1$. To compute the inclusion probability for A,

$$\Pr(i \in A) = \Pr(i \in A_1) + \Pr\left(i \in A_2 | A_1^c\right) \Pr\left(i \in A_1^c\right).$$

Thus, we can use $\pi_i = \pi_{1i} + (1 - \pi_{1i})\pi_{2i|1}$ to compute the Horvitz-Thompson estimator of the form

$$\hat{Y}_{r,\mathrm{HT}} = \sum_{i \in A} \frac{1}{\pi_i} y_i. \tag{A.1}$$

Note that, instead of (A.1), we can consider the following class of estimators:

$$\hat{Y}_w = W \sum_{i \in A_1} \frac{1}{\pi_{1i}} y_i + (1 - W) \sum_{i \in A_2} \frac{1}{\pi_{2i|1}(1 - \pi_{1i})} y_i := W\hat{Y}_1 + (1 - W)\hat{Y}_2. \tag{A.2}$$

Since $\hat{Y}_1$ and $\hat{Y}_2$ are both unbiased for Y, $\hat{Y}_w$ is also unbiased regardless of the choice of W. A reasonable choice of W is $W = n_1/n$.

Under simple random sampling in both designs, the two estimators are equal to $\hat{Y} = N\bar{y}_n$, where $\bar{y}_n$ is the sample mean of y in A. Writing $\bar{y}_1 = n_1^{-1} \sum_{i \in A_1} y_i$ and $\bar{y}_2 = \sum_{i \in A_2} y_i / n_2$, we have

$$\bar{y}_n = W\bar{y}_1 + (1 - W)\bar{y}_2 \tag{A.3}$$

where $W = n_1/n$. Using

$$V(\bar{y}_1) = \left(\frac{1}{n_1} - \frac{1}{N}\right) S_y^2 \tag{A.4}$$

$$V(\bar{y}_2) = \left(\frac{1}{n_2} - \frac{1}{N}\right) S_y^2$$

$$\text{Cov}(\bar{y}_1, \bar{y}_2) = \text{Cov}\left(\bar{y}_1, \bar{y}_1^c\right) = -\frac{n_1}{N - n_1}\left(\frac{1}{n_1} - \frac{1}{N}\right) S_y^2 = -\frac{1}{N} S_y^2,$$

where $\bar{y}_1^c = \sum_{i \in A_1^c} y_i / (N - n_1)$, we have, for $W = n_1 / n$,

$$V(\bar{y}_n) = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2. \tag{A.5}$$

Also,

$$\text{Cov}(\bar{y}_1, \bar{y}_n) = \text{Cov}[\bar{y}_1, W\bar{y}_1 + (1 - W)\bar{y}_2] = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2. \tag{A.6}$$

If $W = n_1 / n$ does not hold, then (A.5) and (A.6) do not hold.

In the KLF application in Section 5, since x and y are measuring the same item, we may assume $S_x^2 = S_y^2 = S_{xy}$ and the variance-covariance matrix of the sampling errors can be smoothed as

$$V(a_h, b_h) = \begin{pmatrix} n_1^{-1} & n^{-1} \\ n^{-1} & n^{-1} \end{pmatrix} S_y^2.$$

References

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

Carroll, R.J., Rupert, D. and Stefanski, L.A. (1995). *Measurement error in nonlinear models*. New York: Chapman & Hall.

Fay, R.E., and Herriot, R.A. (1979). Estimation of income from small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

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

Fuller, W.A. (1991). Small area estimation as a measurement error problem. In *Economic Models, Estimation, and Socioeconomic Systems: Essays in Honor of Karl A. Fox*, (Eds., Tij K. Kaul and Jati K. Sengupta), Elsevier Science Publishers, 333-352.

Fuller, W.A. (2009). *Sampling Statistics*. John Wiley & Sons, Inc., Hoboken, NJ.

Jiang, J., Lahiri, P. and Wan, S. (2002). A unified jackknife theory for empirical best prediction with M-estimation. *Annals of Statistics*, 30, 1782-1810.

Kackar, R.N., and Harville, D.A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association,* 79, 853-862.

Kim, J.K., and Rao, J.N.K. (2012). Combining data from two independent surveys: A model-assisted approach. *Biometrika,* 99, 85-100.

Lohr, S.L., and Prasad, N.G.N. (2003). Small area estimation with auxiliary survey data. *The Canadian Journal of Statistics,* 31, 383-396.

Manzi, G., Spiegelhalter, D.J., Turner, R.M., Flowers, J. and Thompson, S.G. (2011). Modelling bias in combining small area prevalence estimates from multiple surveys. *Journal of the Royal Statistical Society A,* 174, 31-50.

Merkouris, T. (2010). Combining information from multiple surveys by using regression for efficient small domain estimation. *Journal of the Royal Statistical Society B,* 68, 509-521.

Pfeffermann, D. (2002). Small area estimation - New developments and directions. *International Statistical Review,* 70, 125-144.

Quenouille, M.H. (1956). Notes on bias in estimation. *Biometrika,* 43, 353-360.

Raghunathan, T.E., Xie, D., Schenker, N., Parsons, V.I., Davis, W.W., Dodd, K.W. and Feuer, E.J. (2007). Combining information from two surveys to estimate county-level prevalence rates of cancer risk factors and screening. *Journal of the American Statistical Association,* 102, 474-486.

Rao, J.N.K. (2003). *Small Area Estimation*. John Wiley & Sons, Inc., Hoboken, NJ.

Schafer, D.W. (2001). Semiparametric maximum likelihood for measurement error model regression. *Biometrics,* 57, 53-61.

Ybarra, L.M.R., and Lohr, S.L. (2008). Small area estimation when auxiliary information is measured with error. *Biometrika,* 95, 919-931.

Survey Methodology, June 2015
Vol. 41, No. 1, pp. 37-55
Statistics Canada, Catalogue No. 12-001-X

Observed best prediction via nested-error regression with potentially misspecified mean and variance

Jiming Jiang, Thuan Nguyen and J. Sunil Rao[1]

Abstract

We consider the observed best prediction (OBP; Jiang, Nguyen and Rao 2011) for small area estimation under the nested-error regression model, where both the mean and variance functions may be misspecified. We show via a simulation study that the OBP may significantly outperform the empirical best linear unbiased prediction (EBLUP) method not just in the overall mean squared prediction error (MSPE) but also in the area-specific MSPE for every one of the small areas. A bootstrap method is proposed for estimating the design-based area-specific MSPE, which is simple and always produces positive MSPE estimates. The performance of the proposed MSPE estimator is evaluated through a simulation study. An application to the Television School and Family Smoking Prevention and Cessation study is considered.

Key Words: Designe-based MSPE; Heteroscedasticity; Model misspecification; OBP; Small area estimation; TVSFP.

1 Introduction

Observed best prediction (OBP; Jiang, Nguyen and Rao 2011) is a new method for small area estimation (SAE; e.g., Rao 2003). It is motivated by the fact that the best linear unbiased prediction (BLUP) is a hybrid of best prediction (BP) and maximum likelihood (ML) estimation, while the main interest in SAE is typically a prediction problem. The OBP derives parameter estimation based on a purely predictive consideration, leading to the so-called best predictive estimator (BPE) of the model parameters. The development of the OBP in Jiang et al. (2011) mainly focuses on the Fay-Herriot model (Fay and Herriot 1979). Another important class of SAE models is the nested-error regression (NER) model, introduced by Battese, Harter and Fuller (1988). The NER model may be expressed as

$$y_{ij} = x_{ij}'\beta + v_i + e_{ij}, \tag{1.1}$$

$i = 1, \ldots, m,\ j = 1, \ldots, n_i,$ where the v_i's are the area-specific random effects and e_{ij}'s are errors which are assumed to be independent and normally distributed with mean zero, $\text{var}(v_i) = \sigma_v^2$ and $\text{var}(e_{ij}) = \sigma_e^2$, where σ_v^2 and σ_e^2 are unknown. Under the NER model, the small area mean, assuming infinite population, is $\theta_i = \bar{X}_i'\beta + v_i$ for the i^{th} small area, where $\bar{X}_i$ is the population mean of the x_{ij}'s (assumed known; e.g., Rao 2003). It is seen that θ_i is a (linear) mixed effect. Let $\gamma = \sigma_v^2/\sigma_e^2$. Then, the best predictor (BP) of θ_i, is obtained by minimizing the model-based mean squared prediction error (MSPE),

$$\text{E}_M\left(\tilde{\theta}_i - \theta_i\right)^2, \tag{1.2}$$

1. Jiming Jiang, Thuan Nguyen and J. Sunil Rao, University of California, Davis, Oregon Health and Science University and University of Miami. E-mail: jimjiang@ucdavis.edu.

where E_M denotes expectation under the assumed NER model, and $\breve{\theta}_i$ denotes a predictor of θ_i. By normal theory (e.g., Jiang 2007, page 237), the BP is given by

$$\tilde{\theta}_i = \mathrm{E}_M(\theta_i | y_i) = \bar{X}_i'\beta + \frac{n_i\gamma}{1 + n_i\gamma}(\bar{y}_{i\cdot} - \bar{x}_{i\cdot}'\beta), \tag{1.3}$$

where $y_i = (y_{ij})_{1\le j\le n_i}$, β and γ are the true parameters, $\bar{y}_{i\cdot} = n_i^{-1}\sum_{j=1}^{n_i} y_{ij}$ and $\bar{x}_{i\cdot} = n_i^{-1}\sum_{j=1}^{n_i} x_{ij}$. The traditional best linear unbiased prediction (BLUP) method is based on (1.3) with β replaced by its ML estimator, assuming that γ is known; and the empirical BLUP (EBLUP) is derived from the BLUP with γ replaced by a consistent estimator.

The OBP procedure (Jiang et al. 2011) derives estimators of β and γ, namely the BPE, by minimizing the observed, design-based MSPE, which is completely different from the traditional methods such as maximum likelihood (ML) and restricted maximum likelihood (REML; e.g., Jiang 2007). Throughout this paper, we assume that the samples are drawn via simple random sampling, without replacement, from each small area, which is what the design-based approach is based upon. Write $\psi = (\beta', \gamma)'$. Note that, in practice, the small area populations are finite. Following Jiang et al. (2011), we consider a super-population NER model. Suppose that the subpopulations of responses $\{Y_{ik}, k = 1, \ldots, N_i\}$ and auxiliary data $\{X_{ikl}, k = 1, \ldots, N_i\}, l = 1, \ldots, p$ are realizations from corresponding super-populations that are assumed to satisfy the NER model. It follows that

$$Y_{ik} = X_{ik}'\beta + v_i + e_{ik}, \quad i = 1, \ldots, m, \quad k = 1, \ldots, N_i, \tag{1.4}$$

where β, v_i and e_{ik} satisfy the same assumptions as in (1.1). Under the finite-population setting, the true small area mean is $\theta_i = \bar{Y}_i = N_i^{-1}\sum_{k=1}^{N_i} Y_{ik}$ (as opposed to $\theta_i = \bar{X}_i'\beta + v_i$ under the infinite-population setting) for $1 \le i \le m$. Furthermore, write $r_i = n_i / N_i$. Then, the finite-population version of the BP (1.3) has the expression (e.g., Rao 2003, Section 7.2.5)

$$\tilde{\theta}_i = \mathrm{E}_M(\theta_i | y_i) = \bar{X}_i'\beta + \left\{r_i + (1 - r_i)\frac{n_i\gamma}{1 + n_i\gamma}\right\}(\bar{y}_{i\cdot} - \bar{x}_{i\cdot}'\beta), \tag{1.5}$$

where E_M denotes (conditional) expectation under the assumed super-population NER model, and β and γ are the true parameters. Note that the BP is model-dependent.

In practice, any assumed model is subject to misspecification. Jiang et al. (2011) considers misspecification of the mean function, while assuming that the variance-covariance structure of the data is correctly specified. However, the latter, too, may be misspecified in practice. In this paper, we extend the potential model misspecification to both the mean function and the variance-covariance structure. One possible misspecification of the variance-covariance structure is heteroscedasticity, defined in terms of $\mathrm{var}(e_{ij}) = \sigma_i^2$ for area $i, 1 \le i \le m$, where the σ_i^2's are unknown and possibly different. However, in spite of the potential model misspecification, there are reasons that one cannot "abandon" the assumed model, and the model-based BP. First, the assumed model and BP are relatively simple to use, and therefore, attractive to practitioners; in particular, they utilizes simple relationship (linear) between the response and auxiliary variables. For example, in contrast to (1.4), which may subject to misspecification

of the mean function, $X_{ik}'\beta$, one may assume $Y_{ik} = \mu_{ik} + v_i + e_{ik}$, where the μ_{ik} are completely unspecified, unknown constants. The latter model is almost always correct, but is useless, because it does not utilize any relationship between Y and X at all. In fact, in practice, if auxiliary data are available, it is often "politically incorrect" not to use them. Secondly, even though there is a concern about the model misspecification, it often lacks (statistical) evidence on why something else is more reasonable, or whether a complication is necessary. For example, sometimes there is a concern about the normality assumption, but there is no indication on why an alternative distribution, say, t_5, is more reasonable. As another example, suppose that one fits a quadratic model and finds that the coefficient of the quadratic term is insignificant. Then, one is not sure whether the complication of quadratic modeling is necessary as opposed to linear modeling. Thus, as far as this paper is concerned, we are not attempting to change the assumed model, or the BP, (1.5), based on the assumed model. In particular, we assume a single parameter, γ, in (1.5) for the ratio σ_v^2/σ_e^2, rather than considering a heteroscedastic NER model such as in Jiang and Nguyen (2012), and Nandram and Sun (2012). Our goal is to find a better way to estimate the parameters, ψ, under the assumed model that are involved in (1.5), so that the resulting BP, (1.5), is more robust against model misspecifications. We do so by considering an objective MSPE that is not model-dependent, defined as follows. Let $\theta = (\theta_i)_{1\le i\le m}$ denote the vector of small area means, and $\tilde{\theta} = \left[\tilde{\theta}_i\right]_{1\le i\le m}$ the vector of BPs. Note that $\tilde{\theta}_i$ depends on ψ, that is, $\tilde{\theta}_i = \tilde{\theta}_i(\psi)$. The design-based MSPE is

$$\mathrm{MSPE}\left(\tilde{\theta}\right) = \mathrm{E}\left(\left|\tilde{\theta} - \theta\right|^2\right) = \sum_{i=1}^{m} \mathrm{E}\left\{\tilde{\theta}_i(\psi) - \theta_i\right\}^2. \tag{1.6}$$

Note that the E in (1.6) is different from the E_M in (1.2), (1.3), or (1.5) in that E is completely model-free; namely, the expectation in (1.6) is with respect to the simple random sampling from the areas, which has nothing to do with the assumed model. Jiang et al. (2011) showed that the MSPE in (1.6) has an alternative expression, which is a key idea of the OBP. Namely, we have $\mathrm{MSPE}\left(\tilde{\theta}\right) = \mathrm{E}\{Q(\psi) + \cdots\}$, where $\cdots$ does not depend on ψ, and

$$Q(\psi) = \sum_{i=1}^{m}\left\{\tilde{\theta}_i^2(\psi) - 2\frac{1-r_i}{1+n_i\gamma}\bar{y}_{i\cdot}\bar{X}_i'\beta + b_i(\gamma)\hat{\mu}_i^2\right\} = \sum_{i=1}^{m} Q_i. \tag{1.7}$$

In (1.7), ψ is considered as a parameter vector, rather than the true parameter vector, $b_i(\gamma) = 1 - 2a_i(\gamma)$ with $a_i(\gamma) = r_i + (1-r_i)n_i\gamma(1+n_i\gamma)^{-1}$. Furthermore, $\hat{\mu}_i^2$ is a design-unbiased estimator of $\bar{Y}_i^2$ that has the following expression:

$$\hat{\mu}_i^2 = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}^2 - \frac{N_i - 1}{N_i(n_i - 1)}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_{i\cdot}\right)^2. \tag{1.8}$$

The BPE of $\psi, \hat{\psi}$, is the minimizer of $Q(\psi)$ with respect to ψ. For the reader's convenience, the derivations of (1.7) and (1.8) are provided in the Appendix. Also note that the BP is based on the (model-based) area-specific MSPE (so it is optimal for every small area, if the assumed model is correct), while the BPE is based on the (design-based) overall MSPE. This is because we do not want the estimator of ψ to be area-dependent. One reason is that area-dependent estimators are often unstable due to the small

sample size from the area, while an estimator obtained by utilizing all of the areas, such as the BPE defined in this paper, tends to be much more stable.

The consideration of the design-based MSPE, as we do in this paper, is due to the fact that the design-based MSPE is completely model-free. Note that, in Jiang et al. (2011), where the authors considered the Fay-Herriot model, it is not possible to evaluate the design-based MSPE, because the actual samples from the areas are not available (only summaries of the data are available at the area level). Thus, instead, the authors considered model-based MSPE under the most general, or least restrictive, model, which simply assumes that the mean function is μ_i, where μ_i is completely unknown, for the i^{th} small area. In general, there is a "rule of thumb" on what kind of MSPE one should consider. Essentially, the rule is that one should make the MSPE as model-free as possible, so that it would be objective and (relatively) robust to model-misspecifications.

In Section 2, we first consider a simulated example in which we compare the design-based predictive performance of the OBP with that of the EBLUP. Such comparisons were made in Jiang et al. (2011) under the Fay-Herriot model, but has never been done under the NER model. Furthermore, the simulation setting involves misspecification of both the mean function and the variance function, which, again, has not been considered. The simulation results show that the OBP can outperform the EBLUP not just in the overall design-based MSPE but also in the (design-based) area-specific MSPE for every one of a large number of small areas. This is clearly something that has never been discovered. For example, in Jiang et al. (2011), the OBP is shown to outperform the EBLUP in the overall MSPE but not necessarily for every small area.

An important problem of practical interest is estimation of the area-specific MSPEs, here the design-based MSPEs. In Section 3, we propose a bootstrap estimator for the area-specific MSPE, which has the advantage of simplicity and always being positive. Another simulation study is carried out to evaluate the performance of the proposed MSPE estimator. An application to the Television School and Family Smoking Prevention and Cessation Project (TVSFP) is discussed in Section 4.

2 Simulation studies: OBP vs EBLUP

2.1 A demonstration

We first use a simple simulated example to demonstrate the potential impact of model misspecification in terms of the design-based predictive performance of the OBP and the EBLUP. Consider a case where a single covariate, x_{ij}, is thought to be linearly associated with the response y_{ij} through the following NER model:

$$y_{ij} = \beta x_{ij} + v_i + e_{ij}, \quad i = 1, \ldots, m, \quad j = 1, \ldots, 5 \tag{2.1}$$

(so we have $n_i = 5, 1 \le i \le m$ in this case), where β is an unknown coefficient, and v_i, e_{ij} are the same as in (1.1). Thus, in particular, there is a belief that the mean response should be zero when the value of the covariate is zero.

We consider three different sample sizes: $m = 50$, 100 or 400 in conjunction with two different true values of b : $b = 0.5$ or 1.0, where b is defined below. Thus, there are six cases, each being a combination of the sample size and b value. In each case, an x subpopulation is generated from the normal distribution with mean equal to 1 and standard deviation equal to $\sqrt{0.1} \approx 0.32$. The y subpopulation is then generated from the following super-population heteroscedastic NER model:

$$Y_{ik} = b + v_i + e_{ik}, \quad i = 1, \ldots, m, \quad k = 1, \ldots, 1{,}000 \tag{2.2}$$

(so the subpopulation size is $N_i = 1{,}000, 1 \le i \le m$), where v_i is generated from the normal distribution with mean 0 and standard deviation $\sqrt{0.1} \approx 0.32$; e_{ij} is generated from the normal distribution with mean 0 and standard deviation σ_i, where σ_i^2 are generated independently from the Uniform $[0.05, 0.15]$ distribution (so that range for σ_i is approximately from 0.22 to 0.39); and the v_i's and e_{ik}'s are generated independently. It is seen that the assumed NER model is misspecified in terms of both the mean and the variance functions. Once the x and y subpopulations are generated, they are fixed throughout the simulations.

In each simulation, we draw a simple random sample of size 5 from $\{1, \ldots, 1{,}000\}$ that determines the samples x_{ij} and y_{ij}, $j = 1, \ldots, 5$, for each i. This is repeated for $K = 1{,}000$ simulation runs. We make same-data comparisons of the OBP and EBLUP, with the ML estimator of γ for the latter, in terms of both the overall and area-specific MSPEs. The overall MSPE is defined as $\text{MSPE}(\hat{\theta}) = \text{E}\left(|\hat{\theta} - \theta|^2\right) = \sum_{i=1}^{m} \text{E}\left(\hat{\theta}_i - \theta_i\right)^2$, where $\theta = (\theta_i)_{1 \le i \le m}$ is the vector of true small area means with $\theta_i = \bar{Y}_i$, and $\hat{\theta} = \left(\hat{\theta}_i\right)_{1 \le i \le m}$ is the vector of predicted values (either by OBP or by EBLUP). Note that the same measure has been used in Jiang et al. (2011). Table 2.1 reports the overall MSPE results, where the MSPE is evaluated empirically by $K^{-1} \sum_{k=1}^{K} \left|\hat{\theta}^{(k)} - \theta^{(k)}\right|^2 = K^{-1} \sum_{k=1}^{K} \sum_{i=1}^{m} \left\{\hat{\theta}_i^{(k)} - \theta_i^{(k)}\right\}^2$, and $\theta^{(k)} = \left[\theta_i^{(k)}\right]_{1 \le i \le m}$ and $\hat{\theta}^{(k)} = \left[\hat{\theta}_i^{(k)}\right]_{1 \le i \le m}$ are the θ and $\hat{\theta}$ in the k^{th} simulation run, respectively. It is seen that the percentage increase in the overall MSPE of the EBLUP over the OBP ranges between around 20% to almost 1,000%, depending on the sample size and value of b. The patterns shown here are consistent with those in Jiang et al. (2011) under the Fay-Herriot model, where model-based predictive performances are evaluated. However, the gain by the OBP is much more significant, for $m = 100$ and $m = 400$, than those reported in Jiang et al. (2011).

Table 2.1
Overall empirical MSPE (% Increase is EBLUP over OBP)

m	*b*	OBP	EBLUP	% Increase
50	0.5	0.130	0.161	24
50	1.0	0.503	0.598	19
100	0.5	0.076	0.277	264
100	1.0	0.396	1.077	172
400	0.5	0.096	0.965	905
400	1.0	0.393	4.046	930

As for the area-specific MSPEs, following Jiang et al. (2011), we use boxplots to exhibit the distributions of the area-specific MSPEs associated with both methods. See Figure 2.1. The plots reveal details not shown by the overall MSPEs. For example, it might be wondered whether the percentage increase by the EBLUP in the overall MSPE is simply due to the increased number of areas adding together. A simple calculation suggests that this may not be true, for example, $(400/50) \times 19\%$ is only 152% (not 930%). A more explicit explanation is given in Figure 2.1. For example, comparing the case of $m = 50, b = 1$ with that of $m = 400, b = 1$, it is seen that while there is a considerable overlap between the boxplots of OBP and EBLUP in the former case, the boxplots are completely separated in the latter case; in other words, the largest area-specific MSPE of the OBP is smaller than the smallest area-specific MSPE of the EBLUP. This pattern cannot be simply credited to adding or duplicating the areas. In fact, in the latter case, the OBP is doing much better than the EBLUP not just overall, but also for every one of the 400 small areas. This is clearly something never reported before. For example, in the first simulated example of Jiang et al. (2011), the authors found that the OBP has smaller MSPE compared to the EBLUP for half of the small areas while the EBLUP has smaller MSPE compared to the OBP for the other half; similar patterns were found in the second simulated examples in Jiang et al. (2011).

The estimation of the area-specific MSPEs of the OBP is considered in Section 3.

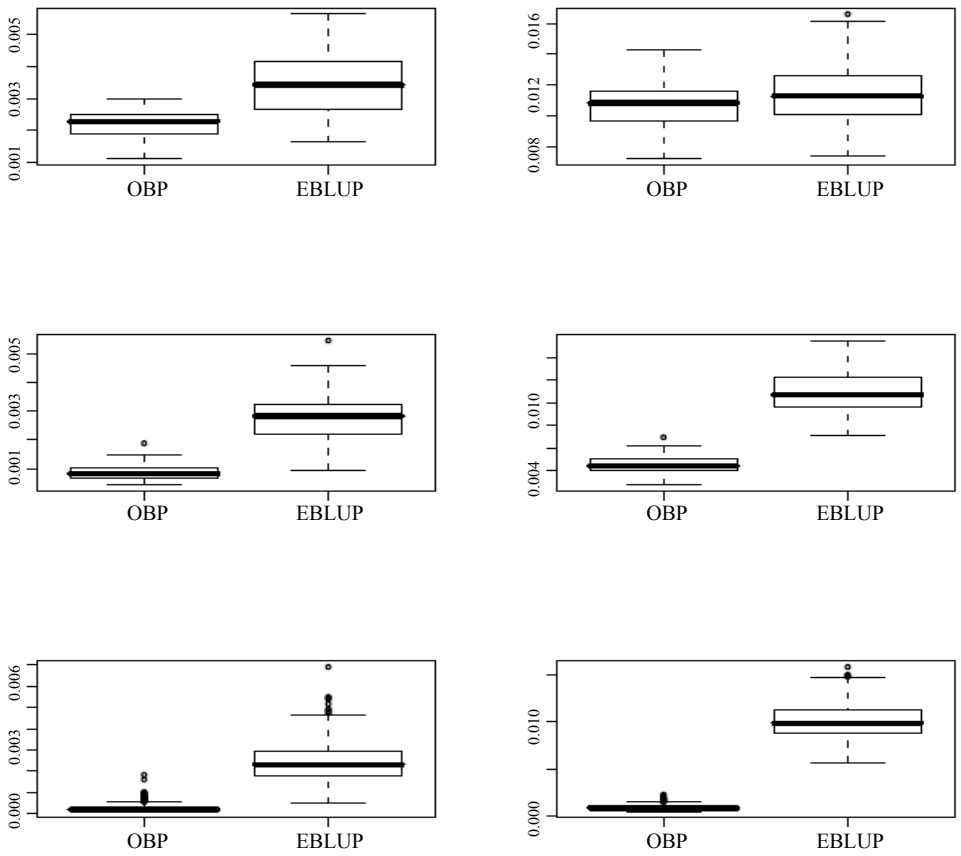


Figure 2.1 Area-specific Empirical MSPEs (Boxplots). Upper Left: $m = 50, b = 0.5$; Upper Right: $m = 50, b = 1.0$; Middle Left: $m = 100, b = 0.5$; Middle Right: $m = 100, b = 1.0$; Lower Left: $m = 400, b = 0.5$; Lower Right: $m = 400, b = 1.0$.

2.2 Further considerations

The situation considered in Subseciton 2.1 might be a little extreme (and this is why we call it a "theoretical demonstration"). In practice, the assumed model may not be completely wrong, or may be close to be correct. In this subsection we first consider a case where the assumed model is "partially correct". Namely, the slope in (2.1) is nonzero (so the assumed model is correct in this regard); the intercept is nonzero, but its value is much smaller compared to those considered in Subsection 2.1 (so the assumed model is wrong, but not "terribly wrong"). More specifically, the true underlying model is

$$Y_{ij} = b_0 + b_1 X_{ik} + v_i + e_{ik}, \quad i = 1, \ldots, m, \quad k = 1, \ldots, 1{,}000, \tag{2.3}$$

as opposed to (2.2), where $b_0 = 0.2, b_1 = 0.1$; the v_i are generated independently from the normal distribution with mean 0 and standard deviation 0.1; and e_{ik} are generated from the heteroscedastic normal distribution as in Subseciton 2.1. In addition to the overall MSPE, we also report contribution to the MSPE due to "bias" and "variance". Let $d_i = \hat{\theta}_i - \theta_i$, and $d_i^{(k)}$ be d_i based on the k^{th} simulated data set, $1 \le k \le K$. We define the empirical bias and variance for the i^{th} small area as $\bar{d}_i = K^{-1} \sum_{k=1}^{K} d_i^{(k)}$ and $v_i^2 = (K-1)^{-1} \sum_{k=1}^{K} \{d_i^{(k)} - \bar{d}_i\}^2$, respectively. Let MSPE_i denote the empirical MSPE for the i^{th} small area. It is easy to show that the overall empirical MSPE is

$$\sum_{i=1}^{m} \text{MSPE}_i = \frac{K-1}{K} \sum_{i=1}^{m} v_i^2 + \sum_{i=1}^{m} (\bar{d}_i)^2 .$$

Thus, the bias and variance contribution to the overall MSPE are defined as $\sum_{i=1}^{m} (\bar{d}_i)^2$ and $\sum_{i=1}^{m} v_i^2$, respectively. Results based on $K = 1{,}000$ simulation runs are presented in Table 2.2. As we can see, for the smaller $m, m = 50,$ OBP performs (slightly) worse than the EBLUP, but for the larger $m, m = 100$ and $m = 400,$ OBP performs (slightly) better, and its advantage increases with $m.$ As for the bias, variance contribution, OBP seems to have smaller bias, and smaller variance for larger $m (m = 100, 400)$.

Table 2.2
Overall Empirical MSPE (bias, variance contribution): Assumed model is partially correct; % Increase is MSPE of EBLUP over MSPE of OBP (negative number indicates decrease)

m	OBP	EBLUP	% Increase
50	0.421 (0.224, 0.197)	0.405 (0.238, 0.167)	-4.0
100	0.733 (0.448, 0.285)	0.748 (0.457, 0.291)	2.1
400	2.745 (1.847, 0.899)	2.848 (1.878, 0.971)	3.8

Next, we consider a case where the assumed model is actually correct. Namely, the true underlying model is (2.3) with $b_0 = 0;$ the errors e_{ik} are homoscedastic with variance equal to 0.1, and everything else is the same as the case considered above. Results based on $K = 1{,}000$ simulation runs are presented in Table 2.3. This time, we see that the EBLUP performs slightly better than OBP under different $m,$ but the difference is diminishing as the sample size increases. As for the bias, variance contribution, EBLUP

seems to have smaller variance, and smaller bias for larger $m\,(m = 100, 400)$, but its advantages in both bias and variance shrink as m increases.

Table 2.3
Overall Empirical MSPE (bias, variance contribution): Assumed model is correct; % Increase is MSPE of EBLUP over MSPE of OBP (negative number indicates decrease)

m	OBP	EBLUP	% Increase
50	0.335 (0.204, 0.131)	0.330 (0.205, 0.125)	-1.4
100	0.749 (0.457, 0.292)	0.746 (0.456, 0.290)	-0.4
400	2.796 (1.800, 0.997)	2.794 (1.799, 0.996)	-0.1

In summary, the simulation results suggest that, when the assumed model is slightly misspecified, OBP may not outperform EBLUP when m, the number of small areas, is relatively small; however, OBP is expected to outperform EBLUP when m is relatively large, and the advantage of OBP over EBLUP increases with m (recall the definition of the overall MSPE). On the other hand, when the assumed model is correct, EBLUP is expected to perform better than OBP, although the difference may be ignorable; and the advantage of EBLUP over OBP is disappearing as m increases. These findings, along with those in Subsection 2.1, are very much in line with those of Jiang et al. (2011; Section 4) under the Fay-Herriot model.

3 Estimation of area-specific MSPE

The design-based area-specific MSPE is defined as

$$\mathrm{MSPE}\left(\hat{\theta}_i\right) = \mathrm{E}\left(\hat{\theta}_i - \theta_i\right)^2, \tag{3.1}$$

where and hereafter E represents the design-based expectation, and $\hat{\theta}_i$ is the OBP of θ_i, given by (1.5) with $\psi = (\beta', \gamma)'$ replaced by its BPE, $\hat{\psi} = \left(\hat{\beta}', \hat{\gamma}\right)'$. As noted in Jiang et al. (2011), it is difficult to obtain second-order unbiased area-specific MSPE estimator under potential model misspecification. This is because the standard asymptotic techniques used in this area, such as the Prasad-Rao linearization method (Prasad and Rao 1990), and the jackknife method (Jiang, Lahiri and Wan 2002), are no longer applicable when the underlying model is misspecified. Jiang et al. (2011) used a different technique to derive a linearization MSPE estimator which is second-order unbiased. However, the latter is not guaranteed nonnegative. Furthermore, the leading term of this MSPE estimator is an $O(1)$ function of the area-specific data, rather than all the data. More precisely, the leading term for the estimated MSPE of $\hat{\theta}_i$, where $\hat{\theta}_i$ is the OBP of the i^{th} small area mean, θ_i, is $\left(\hat{\theta}_i - y_i\right)^2 + D_i\left(2\hat{B}_i - 1\right)$, under the Fay-Herriot model, where y_i is the observation from the i^{th} area (the direct estimator), D_i is the (known) sampling variance, $\hat{B}_i = \hat{A}/\left(\hat{A} + D_i\right)$, and $\hat{A}$ is the BPE for the variance of the area-specific random effect. This is the leading term because its order is $O(1)$, while the rest of the terms in the expression of the estimated

MSPE are of the order $O\left(m^{-1}\right)$ or lower. Because y_i is an observation from a single small area, it has a relatively large variance, that is, the variance is $O(1)$, if n_i is bounded. On the other hand, the BPE $\hat{A}$ is obtained using data from all the small areas, and therefore has a relatively (much) smaller variance; and $\hat{\theta}_i$ is a mixture of y_i and the BPEs. As a result, $\left(\hat{\theta}_i - y_i\right)^2$ is the main contributor to the variance, which can be quite large due to the variation of y_i. On the other hand, the term $D_i\left(2\hat{B}_i - 1\right)$ can be negative. Thus, as a result of the high variation of $\left(\hat{\theta}_i - y_i\right)^2$, there is a non-vanishing probability (as m increases) that the leading term, hence the estimated MSPE, is negative. If we are to take a similar linearization approach under the NER model, we can derive a second-order unbiased MSPE estimator that involves $\bar{y}_{i\cdot}$ in its leading term, which is based on data from a single small area. Then, once again, we run into the problem of large variation and non-vanishing probability of negative value for the MSPE estimator.

Jiang et al. (2011) also used a parametric bootstrap method to obtain an alternative MSPE estimator; however, the justification for the use of this method is questionable given the potential model misspecification. Here we propose to use the nonparametric bootstrap following Efron's original idea (Efron 1979). The method does not rely on the NER model, hence is not affected by the model misspecification. Therefore, the current method is better justified. Furthermore, the proposed MSPE estimator is guaranteed nonnegative, and positive with probability one, which is a major advantage over the linearization MSPE estimator of Jiang et al. (2011).

Suppose that the small area subpopulations, or the N_i's, are large enough, so that the sampling from the subpopulations can be treated approximately as with replacement. Let $z_{ij} = \left(x_{ij}', y_{ij}\right)', j = 1, \ldots, n_i$ denote the (original) samples from the i^{th} small area, $1 \le i \le m$. We then draw samples, $z_{ij}^{(a)} = \left[\left\{x_{ij}^{(a)}\right\}', y_{ij}^{(a)}\right]', j = 1, \ldots, n_i$, with replacement, from $\left\{z_{ij}, j = 1, \ldots, n_i\right\}$, independently for $1 \le i \le m$. Suppose that B bootstrap samples are drawn, yielding samples $z^{(a)} = \left\{z_{ij}^{(a)}, 1 \le j \le n_i, 1 \le i \le m\right\}, 1 \le a \le B$. The bootstrapped version of the BP (1.5) is

$$\tilde{\theta}_i^{(a)} = \bar{X}_{i\cdot}'\beta + \left\{r_i + (1 - r_i)\frac{n_i\gamma}{1 + n_i\gamma}\right\}\left[\bar{y}_{i\cdot}^{(a)} - \left\{\bar{x}_{i\cdot}^{(a)}\right\}'\beta\right], \tag{3.2}$$

where β and γ are the same population parameters for β and γ, respectively, as for the original population. Note that the original samples of z_{ij} are assumed to satisfy the same NER model, (1.4), with $X_{ik}\left(Y_{ik}\right)$ replaced by $x_{ij}\left(y_{ij}\right)$. Because the original samples are treated as the bootstrap population, following Efron's original idea, the population parameters, β, γ, for the bootstrap samples are the same as those for the original samples. Nevertheless, as mentioned, the proposed bootstrap procedure is nonparametric in the sense that the assumed model, (1.4), plays no role in drawing the bootstrap samples. In particular, the BPE of β and γ, based on the original samples, are not used anywhere in the bootstrapping; and the population quantities of interest are $\bar{Y}_i, 1 \le i \le m$, whose bootstrap analogies are $\bar{y}_{i\cdot}, 1 \le i \le m$. This is different from the parametric bootstrap of Jiang et al. (2011), where the BPE of the model parameters, based on the original samples, are used to draw bootstrap samples under the assumed model. Also note that, because the $\bar{X}_i$ are known, they are treated as known constants, and therefore do

not change during the bootstrapping (it does not make sense to "estimate" something that one already knows). Other than those, the procedure follows closely the standard bootstrap idea (e.g., Efron and Tibshirani (1993); also see Chatterjee, Lahiri and Li (2008) for an application to small area estimation). The bootstrap estimator of $\mathrm{MSPE}\left(\hat{\theta}_i\right) = \mathrm{E}\left(\hat{\theta}_i - \bar{Y}_i\right)^2$ is

$$\widehat{\mathrm{MSPE}}\left(\hat{\theta}_i\right) = \frac{1}{B}\sum_{a=1}^{B}\left\{\hat{\theta}_i^{(a)} - \bar{y}_{i\cdot}\right\}^2, \tag{3.3}$$

where $\hat{\theta}_i^{(a)}$ is (3.2) with β, γ replaced by their BPE based on the bootstrapped samples.

Note. One might be concerned that, because the n_i's may be small in typical SAE problems, there may not be many distinct bootstrap samples for each small area. However, the data consist of not just one, but many small areas. When all of the small areas are combined, there are, still, a lot of distinct bootstrap samples, even if the n_i's are small.

We evaluate the performance of the proposed MSPE estimator by considering the simulated example of Subection 2.1 with $b = 0.5$ but under smaller sample sizes. Namely, we start with the basic sample size $m = 10$ and $n_i = 5$, and then either increase n_i, from 5 to 10, or increase m, from 10 to 20. We first consider the design-based bias of $\widehat{\mathrm{MSPE}}\left(\hat{\theta}_i\right)$. Two finite populations are generated, and then fixed, so that the finite population for $m = 10$ is a subpopulation of the finite population for $m = 20$. Table 3.1 reports, for the first ten small areas (these are all the small areas that are common under different values of m), the simulated true MSPE (MSPE), obtained the same way as in Section 2, the simulated mean of $\widehat{\mathrm{MSPE}}\left(\hat{\theta}_i\right)\left(\widehat{\mathrm{MSPE}}\right)$, and the percentage relative bias (%RB) defined as

$$100 \times \left\{\frac{\mathrm{E}\left(\widehat{\mathrm{MSPE}}\right) - \text{True MSPE}}{\text{True MSPE}}\right\},$$

where the expectation is based on the simulations. Another measure of performance is the square root of the mean squared error (RMSE) over the simulations, defined as

$$\sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(\widehat{\mathrm{MSPE}}_{i,k} - \mathrm{MSPE}_i\right)^2}$$

for the i^{th} small area, where MSPE_i is the true MSPE for the i^{th} small area (which does not depend on k), evaluated over the simulations, and $\widehat{\mathrm{MSPE}}_{i,k}$ is the MSPE estimate based on the k^{th} simulated data set. We consider $B = 100$ as the number of bootstrap samples used to evaluate the MSPE estimator, (3.3). All results are based on 1,000 simulation runs. It is seen that, overall, the results improve when either n_i or m increase, but, in terms of %RB, the improvement is more universal, or effective, when n_i increases. This is mainly due to the fact that, as n_i increases, the sample provides a better approximation to the population; hence, the bootstrap distribution better approximates to the population distribution. Also note that, depending on the area, the sign of the RB can be either positive or negative. This is mainly due to the area-to-area difference (recall that the populations are fixed) as well as the bootstrap errors. To

obtain some overall measures, we report the mean and standard deviation (s.d.) of the %RBs over the ten small areas as follows: $m = 10, n_i = 5$: mean = 4.2%, s.d. = 14.8%; $m = 10, n_i = 10$: mean = 1.5%, s.d. = 4.2%; $m = 20, n_i = 5$: mean = −0.6%, s.d. = 8.1%. The boxplots of the %RBs are presented in Figure 3.1. The plots further illustrate the pattern of improvement. On the other hand, in terms of RMSE, the improvement is much more significant when m increases than when n_i increases. This is because having a larger m reduces the MSPEs, in general; hence, natually, the correponding MSPE estimates also drop. In other words, both the estimator and the parameter (the MSPE) decrease, which typically results in a reduction in RMSE. The summary and boxplots for RMSE are omitted.

In addition, the %RB and RMSE in Table 3.1 fluctuates quite a bit from area to area. This is mainly due to the area to area difference. Recall the small area populations are generated each with population size $N_i = 1{,}000$, and then fixed throughout the simulation. Although the superpopulation used to generate the small area populations, including X and Y, are the same, there are still some differences in the generated finite populations, especially because the population size, N_i, is not very large.

Table 3.1
Empirical Performance of $\widehat{\text{MSPE}}$

m	n_i	i	MSPE	$\widehat{\text{MSPE}}$	%RB	RMSE	i	MSPE	$\widehat{\text{MSPE}}$	%RB	RMSE
10	5	1	0.041	0.042	4.5	0.103	6	0.034	0.043	26.3	0.070
10	10	1	0.036	0.036	-0.4	0.068	6	0.034	0.036	6.4	0.070
20	5	1	0.031	0.032	4.1	0.051	6	0.028	0.031	12.5	0.046
10	5	2	0.046	0.038	-16.1	0.078	7	0.032	0.040	25.4	0.078
10	10	2	0.035	0.033	-4.1	0.078	7	0.033	0.034	2.7	0.068
20	5	2	0.031	0.029	-7.2	0.050	7	0.030	0.031	3.6	0.055
10	5	3	0.038	0.042	10.2	0.121	8	0.042	0.042	-0.4	0.150
10	10	3	0.037	0.036	-1.7	0.091	8	0.033	0.035	7.5	0.067
20	5	3	0.031	0.032	4.4	0.052	8	0.030	0.031	4.1	0.058
10	5	4	0.056	0.052	-7.6	0.121	9	0.050	0.042	-15.0	0.074
10	10	4	0.037	0.040	6.3	0.072	9	0.034	0.034	-1.0	0.063
20	5	4	0.040	0.035	-11.3	0.068	9	0.034	0.030	-11.1	0.049
10	5	5	0.033	0.037	11.8	0.066	10	0.041	0.043	3.1	0.082
10	10	5	0.032	0.033	2.5	0.066	10	0.034	0.033	-2.9	0.073
20	5	5	0.024	0.025	2.9	0.052	10	0.035	0.033	-7.9	0.062

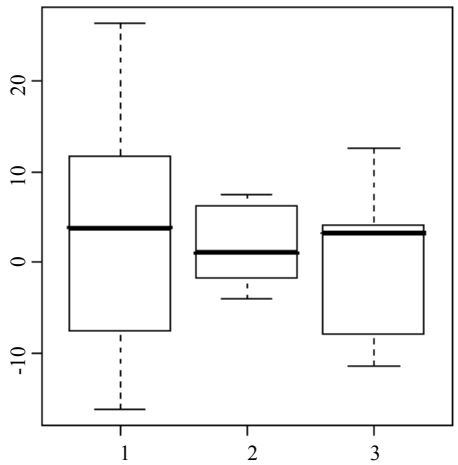


Figure 3.1 Boxplots of %RB. 1 : $m = 10, n_i = 5$; 2 : $m = 10, n_i = 10$; 3 : $m = 20, n_i = 5$.

We conclude this section with some comments on the theoretical side. While there have been extensive studies on MSPE estimation in SAE since Prasad and Rao's seminal paper (Prasad and Rao 1990), the vast majority of these work focus on the model-based MSPEs. See, for example, Datta, Kubokawa, Molina and Rao (2011), Lahiri (2012), and Torabi and Rao (2012) for some recent work on design-based MSPE estimation in SAE. As noted in Jiang et al. (2011), under possible model misspecification, the model-based area-specific MSPE is not consistently estimable, and this is true for the design-based, area-specific MSPE as well. The reason is that, when the model is misspecified in terms of the mean function, the MSPE is not a function of a finite number of parameters (such as $\beta, \gamma,$ and σ_e^2). In fact, because we operate under possible model misspecification, the quantities such as $\bar{Y}_i^2, 1 \le i \le m$ are involved in the expressions of the area-specific MSPEs, which should all be treated as unknown parameters. Furthermore, the effective sample size for estimating $\bar{Y}_i^2$ is n_i, if the assumed model fails. It follows that $\bar{Y}_i^2$ cannot be consistently estimated using data from the area alone, if n_i is bounded. Generally speaking, if the MSPE can be estimated consistently, the difference between the MSPE estimator and the MSPE is $O_{\mathrm{P}}\left(m^{-1/2}\right)$; therefore, the bias is typically $O\left(m^{-1}\right)$ without bias correction. On the other hand, if the (area-specific) MSPE cannot be consistently estimated, the difference between the MSPE estimator and the MSPE is typically $O_{\mathrm{P}}\left\{(m \wedge n_i)^{-1/2}\right\}$, where $m \wedge n = \min(m, n)$, hence the bias is typically $O\left\{(m \wedge n_i)^{-1}\right\}$, without the bias correction. The bootstrap MSPE estimator, $\widehat{\mathrm{MSPE}}$, has the latter property, plus that it is always nonnegative. Although it is possible to bias-correct $\widehat{\mathrm{MSPE}}$ to reduce the order of the bias to $o\left\{(m \wedge n_i)^{-1}\right\}$ (e.g., Hall and Maiti 2006), the nonnegative property may be lost after the bias correction. In view of the above discussion, it seems that, under the potential model misspecification, it is reasonable to define the first and second-order unbiasedness of an area-specific MSPE estimator in terms of $O\left\{(m \wedge n_i)^{-1}\right\}$ and $o\left\{(m \wedge n_i)^{-1}\right\}$, instead of the traditional $O\left(m^{-1}\right)$ and $o\left(m^{-1}\right)$ (e.g., Rao 2003).

4 An application

We consider an application of the methods developed in the previous sections to the TVSFP data. For a complete description of the TVSFP study, see Hedeker, Gibbons and Flay (1994). The original study was designed to test independent and combined effects of a school-based social-resistance curriculum and a television-based program in terms of tobacco use prevention and cessation. The subjects were seventh-grade students from Los Angeles (LA) and San Diego in the State of California in the United States. The students were pretested in January 1986 in an initial study. The same students completed an immediate postintervention questionnaire in April 1986, a one-year follow-up questionnaire (in April 1987), and a two-year follow-up (in April 1988). In this analysis, we consider a subset of the TVSFP data involving students from 28 LA schools, where the schools were randomized to one of four study conditions: (a) a social-resistance classroom curriculum (CC); (b) a media (television) intervention (TV); (c) a combination of CC and TV conditions; and (d) a no-treatment control. A tobacco and health knowledge scale (THKS) score was one of the primary study outcome variables, and the one used for this analysis. The THKS consisted of seven questionnaire items used to assess student tobacco and health knowledge. A student's

THKS score was defined as the sum of the items that the student answered correctly. Only data from the pretest and immediate postintervention are available for the current analysis. More specifically, the data only involved subjects who had completed the THKS at both of these time points. On the one hand, the Complete-record data set up an ideal "before-after" situation; on the other hand, the missing data, that is, those from subjects who had completed the questionnaire at only one time point, might have provided additional useful information. For example, it is possible that a subject did not complete the follow-up because he or she did not find the program helpful. Unfortunately, the incomplete data were not available. As a result, there is a potential risk of selection bias for the complete-record-only analysis. In all, there were 1,600 students from the 28 schools, with the number of students from each school ranging from 18 to 137.

Hedeker et al. (1994) carried out a mixed-model analysis based on a number of NER models to illustrate maximum likelihood estimation for the analysis of clustered data. Here we consider a problem of estimating the small area means for the difference between the immediate postintervention and pretest THKS scores (the response). Here the "small area" is understood as a number of major characteristics (e.g., residential area, teacher/student ratio) that affect the response, but are not captured by the covariates in the model (i.e., linear combination of the CC, TV and CCTV indicators). Note that, traditionally, the words "small areas" correspond to small geographical regions or subpopulations, for which adequate samples are not available (e.g., Rao 2003), and such information as residential characteristics or teacher/student ratios would be used as additional covariates. However, such characteristic information are not available. This is why we define these unavailable information as "area-specific", so that they can be treated as the (small-area) random effects. This is consistent with the fundamental features of the random effects that are often used to capture unobservable effects or information (e.g., Jiang 2007), and extends the traditional notion of small area estimation. Thus, a small area is the seventh graders in all of the U.S. schools that share the similar major characteristics as a LA school involved in the data over a reasonable period of time (e.g., five years) so that these characteristics had not changed much during the time and neither had the social/educational relevance of the CC and TV programs. There are 28 LA schools in the TVSFP data that correspond to 28 sets of characteristics, so that the data are considered random samples from the 28 small areas defined as above. As such, each small area population is large enough so that $n_i/N_i \approx 0, 1 \le i \le 28$. Recall that the n_i's in the TVSFP sample range from 18 to 137, while the N_i's are expected to be at least tens of thousands. Note that the only place in the OBP where the knowledge of N_i is required is through the ratio n_i/N_i. The proposed NER model can be expressed as (1.1) with $x_{ij}'\beta = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,1}x_{i,2}$, where $x_{i,1} = 1$ if CC, and 0 otherwise; $x_{i,2} = 1$ if TV, and 0 otherwise. It follows that all the auxiliary data x_i are at the area level; as a result, the value of $\bar{X}_i$ is known for every i.

As noted, the sample sizes for some small areas are quite large, but there are also areas with relatively (much) smaller sample sizes. This is quite common in real-life problems. Because the auxiliary data are at area-level, we have $\bar{X}_i'\beta = \bar{x}_i'\beta$; thus, it is easy to show that the BP (1.5) can be expressed as

$$\tilde{\theta}_i = \left\{ r_i + (1 - r_i)\frac{n_i\gamma}{1 + n_i\gamma} \right\} \bar{y}_i + \frac{1 - r_i}{1 + n_i\gamma}\bar{x}_i'\beta.$$

It is seen that, when n_i is large, the BP is approximately equal to $\bar{y}_i$, the design-based estimator, which has nothing to do with the parameter estimation. Therefore, when n_i is large, there is not much difference

between the OBP and the EBLUP. On the other hand, when n_i is small or moderate, we expect some difference between the OBP and the EBLUP in terms of the MSPE. However, it is difficult to tell how much difference there is in this real data example. Our simulation results in Section 2 show that the difference between OBP and EBLUP in terms of the MSPE depends on to what extent the assumed model is misspecified. It should be noted that the response, y_{ij}, is difference in the THKS scores, and possible values of the THKS score are integers between 0 and 7. Clearly, such data is not normal. The potential impact of the nonnormality is two-fold. On the one hand, it is likely that the NER model, as proposed by Hedeker et al. (1994), is misspecified, in which case expression (1.5) is no longer the BP, and the Gaussian ML (REML) estimators are no longer the true ML (REML) estimators. On the other hand, even without the normality, (1.5) can still be justified as the best linear predictor (BLP; e.g., Searle, Casella and McCulloch 1992, Section 7.3). Furthermore, the Gaussian ML (REML) estimators are consistent and asymptotically normal even without the normality assumption (Jiang 1996; also see Jiang 2007, Chapter 1). Other aspects of the NER model include homoscedasticity of the error variance across the small areas. Figure 4.1 shows the histogram of the sample variances of the 28 small areas. The bimodal shape of the histogram suggests potential heteroscedasticity in the error variance, yet another type of possible model misspecification. Therefore, the OBP method is naturally considered.

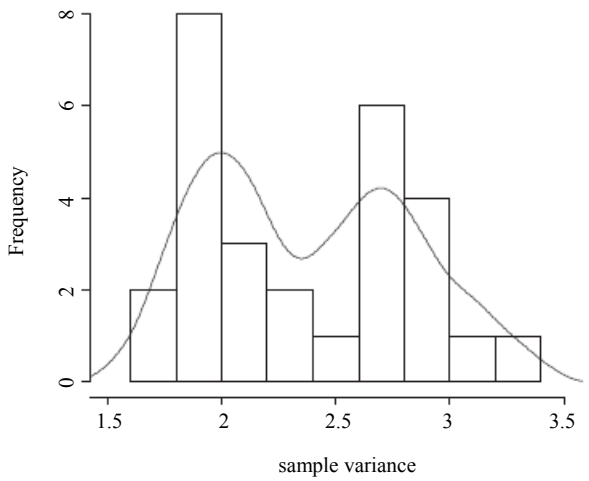


Figure 4.1 Histogram of sample variances; a kernel density smoother is fitted.

We carry out the OBP analysis for the 28 small areas and the results are presented in Table 4.1. The BPE of the parameters are $\hat{\beta}_0 = 0.206, \hat{\beta}_1 = 0.687, \hat{\beta}_2 = 0.213, \hat{\beta}_3 = -0.288$, and $\hat{\gamma} = 0.003$. Although interpretation may be given for the parameter estimates, there is a concern about possible model misspecification (in which case the interpretation may not be sensible), as noted earlier. Regardless, our main interest is prediction, not estimation; thus, we focus on the OBP. In addition to the OBPs, we also computed the corresponding $\widehat{\text{MSPE}}$, and their square roots as the measures of uncertainty. As a comparison, the EBLUPs for the small areas as well as the corresponding square roots of the MSPE estimates, $\widetilde{\text{MSPE}}$, using the Prasad-Rao method (P–R; Prasad and Rao 1990) are also included in the table. It is seen that the OBPs are all positive, even for the small areas in the control group. As for the

statistical significance (here "significance" is defined as that the OBP is greater in absolute value than 2 times the corresponding square root of the MSPE estimate), the small area means are significantly positive for all of the small areas in the (1,1) group. In contrast, none of the small area mean is significantly positive for the small areas in the (0,0) group. As for the other two groups, the small area means are significantly positive for all the small areas in the (1,0) group; the small area means are significantly positive for all but two small areas in the (0,1) group. There are 7, 8, 7 and 7 small areas in the (0,0), (0,1), (1,0) and (1,1) groups, respectively.

Table 4.1
OBP, EBLUP, measures of uncertainty for TVSFP data (Part 1)

ID	CC	TV	OBP	$\sqrt{\widehat{MSPE}}$	EBLUP	$\sqrt{\widehat{MSPE}}$
403	1	0	0.886	0.171	0.913	0.121
404	1	1	0.844	0.296	0.856	0.121
193	0	0	0.215	0.207	0.217	0.120
194	0	0	0.221	0.137	0.221	0.134
196	1	0	0.878	0.171	0.907	0.124
197	0	0	0.225	0.158	0.223	0.126
198	1	1	0.771	0.220	0.807	0.131
199	0	1	0.426	0.142	0.453	0.130
401	1	1	0.826	0.133	0.844	0.127
402	0	0	0.188	0.171	0.199	0.123
405	0	1	0.394	0.147	0.432	0.129
407	0	1	0.508	0.300	0.508	0.133
408	1	0	0.871	0.240	0.903	0.123
409	0	0	0.230	0.125	0.227	0.136

Table 4.2
OBP, EBLUP, measures of uncertainty for TVSFP data (Part 2)

ID	CC	TV	OBP	$\sqrt{\widehat{MSPE}}$	EBLUP	$\sqrt{\widehat{MSPE}}$
410	1	1	0.778	0.304	0.813	0.124
411	0	1	0.409	0.195	0.444	0.115
412	1	0	0.913	0.219	0.930	0.126
414	1	0	0.929	0.257	0.941	0.127
415	1	1	0.869	0.199	0.872	0.135
505	1	1	0.790	0.154	0.818	0.136
506	0	1	0.389	0.169	0.428	0.134
507	0	1	0.426	0.148	0.452	0.135
508	0	1	0.411	0.108	0.442	0.136
509	1	0	0.915	0.097	0.929	0.143
510	1	0	0.880	0.119	0.905	0.143
513	0	0	0.185	0.215	0.197	0.123
514	1	1	0.866	0.144	0.870	0.140
515	0	0	0.180	0.102	0.192	0.143

Comparing the OBP with the EBLUP, the values of the latter are generally higher, and their corresponding MSPE estimates are mostly lower. In terms of statistical significance, the EBLUP results are significant for the (1,1), (1,0) and (0,1) groups, and insignificant for the (0,0) group. It should be noted that the $P-R$ MSPE estimator for the EBLUP is derived under the normality assumption, while in this

case the data is clearly not normal, as noted earlier. Thus, the measure of uncertainty for the EBLUP may not be accurate. In particular, just because the (square roots of the) MSPEs for the EBLUPs are lower, compared to those for the OBPs, it does not mean the corresponding true MSPEs for the EBLUPs are lower than those for the OBPs. In fact, our simulation results (see Section 2) have shown otherwise. It is also observed that the MSPE estimates for the EBLUPs are more homogeneous cross the small areas. This may be due to the fact that the $\mathrm{P}-\mathrm{R}$ MSPE estimator for EBLUP is obtained assuming that the NER model is correct, while the proposed MSPE estimator for OBP does not use such an assumption.

In conclusion, in spite of the potential difference in the small area characteristics, the CC and TV programs appear to be successful in terms of improving the students' THKS scores (whether the improved THKS score means improved tobacco use prevention and cessation is a different matter though). It also seems apparent that the CC program was relatively more effective than the TV program. Without the intervention of any of these programs, the THKS score did not seem to improve in terms of the small area means. In terms of the statistically significant results, when CC = 0 and TV = 0, the THKS score did not seem to improve; when CC = 1, the THKS score seemed to improve; and, when CC = 0 and TV = 1, the improvement of the THKS score was not so convincing.

Acknowledgements

Jiming Jiang is partially supported by the NSF grants DMS-0809127 and SES-1121794. Thuan Nguyen is partially supported by the NSF grant SES-1118469. J. Sunil Rao is partially supported by the NSF grants DMS-0806076 and SES-1122399. The research of all three authors are partially supported by the NIH grant R01-GM085205A1. The authors thank Professor Donald Hedeker for kindly providing the TVSFP data for our analysis. Finally, the authors are grateful to the comments made by an Associate Editor and two referees.

Appendix

A.1. OBP under nested-error regression. The design-based MSPE is given by (1.6). Note that all the E, and later P, are design-based, assuming simple random sampling. Note that $\mathrm{E}\{\tilde{\theta}_i(\psi)-\theta_i\}^2 = \mathrm{E}\{\tilde{\theta}_i^2(\psi)\} - 2\theta_i\mathrm{E}\{\tilde{\theta}_i(\psi)\} + \theta_i^2$. Furthermore, note that $\mathrm{E}(\bar{y}_{i\cdot}) = \theta_i$ and $\mathrm{E}(\bar{x}_{i\cdot}) = \bar{X}_i$ ($\bar{y}_{i\cdot}$ and $\bar{x}_{i\cdot}$ are design-unbiased estimators of their corresponding subpopulation means). Thus, we have

$$\begin{aligned}\mathrm{E}\{\tilde{\theta}_i(\psi)\} &= \bar{X}_i'\beta + \left\{\frac{n_i}{N_i} + \left(1-\frac{n_i}{N_i}\right)\frac{n_i\sigma_v^2}{\sigma_e^2+n_i\sigma_v^2}\right\}(\theta_i - \bar{X}_i'\beta)\\ &= \left(1-\frac{n_i}{N_i}\right)\frac{\sigma_e^2}{\sigma_e^2+n_i\sigma_v^2}\bar{X}_i'\beta + \left\{\frac{n_i}{N_i} + \left(1-\frac{n_i}{N_i}\right)\frac{n_i\sigma_v^2}{\sigma_e^2+n_i\sigma_v^2}\right\}\theta_i.\end{aligned}$$

Thus, using the notation introduced below (1.7), we have

$$\mathrm{E}\{\tilde{\theta}_i(\psi)-\theta_i\}^2 = \mathrm{E}\{\tilde{\theta}_i^2(\psi)\} - 2\frac{1-r_i}{1+n_i\gamma}\bar{X}_i'\beta\theta_i + b_i(\gamma)\theta_i^2. \tag{A.1}$$

We can express the unknown θ_i in (A.1) by $\mathrm{E}(\bar{y}_{i\cdot})$. We also need a design-based unbiased estimator of θ_i^2, which is given by (1.8). In other words, we have $\theta_i^2 = \mathrm{E}(\hat{\mu}_i^2)$. To show the design-unbiasedness of (1.8), note that

$$\begin{aligned}\mathrm{E}\left(\frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}^2\right) &= \frac{1}{n_i}\mathrm{E}\left\{\sum_{k=1}^{N_i} Y_{ik}^2 1_{(k\in I_i)}\right\} \\ &= \frac{1}{n_i}\sum_{k=1}^{N_i} Y_{ik}^2 \mathrm{P}(k \in I_i) = \frac{1}{N_i}\sum_{k=1}^{N_i} Y_{ik}^2,\end{aligned}$$

where I_i is the set of sampled indexes corresponding to the i^{th} small area. Also, we have

$$\begin{aligned}\mathrm{E}\left\{\frac{N_i-1}{N_i(n_i-1)}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2\right\} &= \frac{N_i-1}{N_i(n_i-1)}\mathrm{E}\left(\sum_{j=1}^{n_i} y_{ij}^2 - n_i\bar{y}_{i\cdot}^2\right) \\ &= \frac{N_i-1}{N_i(n_i-1)}\mathrm{E}\left(\sum_{j=1}^{n_i} y_{ij}^2\right) - \frac{(N_i-1)n_i}{N_i(n_i-1)}\mathrm{E}(\bar{y}_{i\cdot}^2) \\ &= \frac{(N_i-1)n_i}{N_i(n_i-1)}\left\{\frac{1}{N_i}\sum_{k=1}^{N_i} Y_{ik}^2 - \mathrm{E}(\bar{y}_{i\cdot}^2)\right\},\end{aligned}$$

and

$$\begin{aligned}\mathrm{E}(\bar{y}_{i\cdot}^2) &= \frac{1}{n_i^2}\mathrm{E}\left\{\sum_{k=1}^{N_i} Y_{ik}1_{(k\in I_i)}\right\}^2 \\ &= \frac{1}{n_i^2}\sum_{k,l=1}^{N_i} Y_{ik}Y_{il}\mathrm{P}(k\in I_i, l\in I_i) \\ &= \frac{1}{n_i^2}\left\{\sum_{k=1}^{N_i} Y_{ik}^2\frac{n_i}{N_i} + \sum_{k\neq l} Y_{ik}Y_{il}\frac{n_i(n_i-1)}{N_i(N_i-1)}\right\} \\ &= \frac{1}{n_i^2}\left[\frac{n_i}{N_i}\sum_{k=1}^{N_i} Y_{ik}^2 + \frac{n_i(n_i-1)}{N_i(N_i-1)}\left\{\left(\sum_{k=1}^{N_i} Y_{ik}\right)^2 - \sum_{k=1}^{N_i} Y_{ik}^2\right\}\right] \\ &= \frac{1}{n_i^2}\left\{\frac{n_i(N_i-n_i)}{N_i(N_i-1)}\sum_{k=1}^{N_i} Y_{ik}^2 + \frac{N_i n_i(n_i-1)}{N_i-1}\theta_i^2\right\} \\ &= \frac{N_i-n_i}{N_i(N_i-1)n_i}\sum_{k=1}^{N_i} Y_{ik}^2 + \frac{N_i(n_i-1)}{(N_i-1)n_i}\theta_i^2.\end{aligned}$$

Thus, after combining things together, we get

$$\mathrm{E}(\hat{\mu}_i^2) = \left[1 - \frac{(N_i-1)n_i}{N_i(n_i-1)}\left\{1 - \frac{N_i-n_i}{(N_i-1)n_i}\right\}\right]\left(\frac{1}{N_i}\sum_{k=1}^{N_i} Y_{ik}^2\right) + \theta_i^2 = \theta_i^2.$$

It follows that the right side of (A.1) can be expressed as

$$\mathrm{E}\left[\sum_{i=1}^{m}\left\{\tilde{\theta}_i^2(\psi) - 2\frac{1-r_i}{1+n_i\gamma}\bar{\mathbf{X}}_i'\beta\bar{y}_{i\cdot} + b_i(\gamma)\hat{\mu}_i^2\right\}\right].$$

The BPE is obtained by minimizing the expression inside the expectation, which is (1.7).

References

Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association,* 83, 401, 28-36.

Chatterjee, S., Lahiri, P. and Li, H. (2008). Parametric bootstrap approximation to the distribution of EBLUP and related prediction intervals in linear mixed models. *The Annals of Statistics,* 36, 3, 1221-1245.

Datta, G.S., Kubokawa, T., Molina, I. and Rao, J.N.K. (2011). Estimation of mean squared error of model-based small area estimators. *Test,* 20, 367-388.

Efron, B. (1979). Bootstrap method: Another look at the jackknife. *The Annals of Statistics,* 7, 1, 1-26.

Efron, B., and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap,* Chapman & Hall/CRC.

Fay, R.E., and Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association,* 74, 366a, 269-277.

Hall, P., and Maiti, T. (2006). Nonparametric estimation of mean-squared prediction error in nested-error regression models. *The Annals of Statistics,* 34, 4, 1733-1750.

Hedeker, D., Gibbons, R.D. and Flay, B.R. (1994). Random-effects regression models for clustered data with an example from smoking prevention research. *Journal of Consulting and Clinical Psychology,* 62, 4, 757-765.

Jiang, J. (1996). REML estimation: Asymptotic behavior and related topics. *The Annals of Statistics,* 24, 1, 255-286.

Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications,* New York: Springer.

Jiang, J., and Nguyen, T. (2012). Small area estimation via heteroscedastic nested-error regression. *The Canadian Journal of Statistics/La revue canadienne de statistique,* 40, 3, 588-603.

Jiang, J., Lahiri, P. and Wan, S.-M. (2002). A unified jackknife theory for empirical best prediction with M - estimation. *The Annals of Statistics,* 30, 6, 1782-1810.

Jiang, J., Nguyen, T. and Rao, J.S. (2011). Best predictive small area estimation. *Journal of the American Statistical Association,* 106, 494, 732-745.

Lahiri, P. (2012). Estimation of average design-based mean squared error of synthetic small area estimators. Presented at the *40th Annual Meeting of the Statistical Society of Canada*, Guelph, ON.

Nandram, B., and Sun, Y. (2012). A Bayesian model for small area under heterogeneous sampling variances. Technical Report.

Prasad, N.G.N., and Rao, J.N.K. (1990). The estimation of mean squared errors of small area estimators. *Journal of the American Statistical Association,* 85, 409, 163-171.

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

Searle, S.R., Casella, G. and McCulloch, C.E. (1992). *Variance Components,* New York: John Wiley & Sons, Inc.

Torabi, M., and Rao, J.N.K. (2012). Estimation of mean squared error of model-based estimators of small area means under a nested error linear regression model. Technical Report.

Survey Methodology, June 2015
Vol. 41, No. 1, pp. 57-77
Statistics Canada, Catalogue No. 12-001-X

A method of determining the winsorization threshold, with an application to domain estimation

Cyril Favre Martinoz, David Haziza and Jean-François Beaumont[1]

Abstract

In business surveys, it is not unusual to collect economic variables for which the distribution is highly skewed. In this context, winsorization is often used to treat the problem of influential values. This technique requires the determination of a constant that corresponds to the threshold above which large values are reduced. In this paper, we consider a method of determining the constant which involves minimizing the largest estimated conditional bias in the sample. In the context of domain estimation, we also propose a method of ensuring consistency between the domain-level winsorized estimates and the population-level winsorized estimate. The results of two simulation studies suggest that the proposed methods lead to winsorized estimators that have good bias and relative efficiency properties.

Key Words: Conditional bias; Robust estimation; Winsorized estimator; Influential values.

1 Introduction

In business surveys, it is not unusual to collect economic variables for which the distribution is highly skewed. In this context, we often face the problem of influential values in the selected sample. These values are typically very large, and their presence in the sample tends to make classical estimators very unstable.

It is possible to guard against the impact of influential values at the design stage by selecting with certainty the potentially influential units. For example, in business surveys, it is customary to use a stratified simple random sampling without-replacement design containing one or more take-all strata that are usually composed of large units. Unfortunately, it is seldom possible to completely eliminate the problem of influential values at the design stage. The strata in business surveys are usually formed using a geography variable, a size variable (for example, number of employees) and a classification variable (for example, the North American Industry Classification System (NAICS) code). In a survey that collects dozens of variables of interest, it is not unlikely that some of them will have little or no correlation with the stratification variables, which may result in the presence of influential values. This is the case in particular in Statistics Canada's environmental surveys, such as the Agricultural Water Survey, one of whose objectives is to measure the quantity of water used by Canadian farms for irrigation. It turns out that water consumption in a given year has little correlation with the stratification variables, since consumption depends in part on the weather conditions affecting the sampled farms. Another example is the Industrial Water Survey, one of whose objectives is to measure the quantity of water used. In the case of mining companies, the consumption of water for ore extraction is strongly correlated with the geophysical characteristics of the land, which are not taken into account by the stratification variables.

Another problem that leads to influential values in the sample is the presence of stratum jumpers, which arises when the stratification information collected in the field is different from the information in

1. Cyril Favre Martinoz, Laboratoire de Statistique d'Enquête, CREST/ENSAI & IRMAR, Campus de Ker Lann, 35170 Bruz, France; David Haziza, Département de mathématiques et statistique, Université de Montréal, Montréal, Canada, H3C 3J7 and Laboratoire de Statistique d'Enquête, CREST/ENSAI, Campus de Ker Lann, 35170 Bruz, France. E-mail: haziza@dms.umontreal.ca; Jean-François Beaumont, Statistical Research and Innovation Division, Statistics Canada, Ottawa, Canada, K1A 0T6.

the sampling frame. These differences are usually due to errors in the frame (for example, an outdated frame). A stratum jumper is a unit that is not in the stratum that it would have been assigned to if the information in the frame had been accurate. If a unit with a large value is assigned to a take-some stratum, it will have a large value for the variable of interest and possibly a large sampling weight, which will potentially make it very influential. In practice, it is not unusual to have between 5% and 10% stratum jumpers.

Classical estimators (such as the expansion estimator) exhibit (virtually) no bias, but they can be very unstable in the presence of a influential values. Robust estimators are constructed so as to limit the impact of influential values, which leads to estimators that are more stable but potentially biased. The objective is to develop robust estimation procedures whose mean square error is significantly smaller than that of classical estimators when there are influential values in the population but which do not suffer a serious loss of efficiency when there are none. The treatment of influential values usually strikes a trade-off between bias and variance.

Winsorization is a method often used in business surveys to treat influential values. It involves decreasing the value and/or weight of one or more influential units to reduce their impact. Two forms of winsorization are considered: standard winsorization and the winsorization described by Dalén (1987) and Tambay (1988). These methods are described in Section 4. Whichever type is used, winsorization requires the determination of a constant that corresponds to the threshold above which large values are reduced. The choice of this constant is crucial, as a poor choice may lead to winsorized estimators that have a larger mean square error than classical estimators. The problem of choosing the constant has been studied by Kokic and Bell (1994) and Rivest and Hurtubise (1995), among others. In the case of a stratified simple random sampling without-replacement design, these researchers determined the constant that minimizes the estimated mean square error of the winsorized estimators. For repeated surveys, they suggest using historical data collected in previous iterations. Kokic and Bell (1994) determined the optimal value of the constant by setting up a common mean model in each stratum and minimizing the winsorized estimator's mean square error with respect to both the model and the sampling design. Clark (1995) generalized the results obtained by Kokic and Bell (1994) to the case of a ratio estimator and by calculating the mean square error with respect to the model only.

First, we consider a different criterion, which involves finding the constant that minimizes the largest estimated conditional bias in the sample. As we explain in Section 2, the conditional bias associated with a unit is a measure of influence that takes into account the sampling design used. The proposed method has the advantage of being simple to apply in practice. In addition, unlike the methods proposed in the literature, it does not require historical information or a model describing the distribution of the variable of interest in each stratum. Robust estimation based on the conditional bias is presented in Section 3.

In Section 5, we deal with the problem of domain estimation, which is an important problem in practice. We apply a robust method separately in each domain of interest. A population-level estimator can easily be produced by aggregating the robust estimators obtained at the domain level. However, since it is defined as the sum of estimators that are all biased, the aggregate estimator could have a large bias. This point was raised by Rivest and Hidiroglou (2004). We propose a three-step approach: First, apply a robust method separately in each domain of interest to produce initial estimates. Independently, produce an initial robust estimate at the population level. Lastly, using a method similar to calibration (e.g., Deville and Särndal 1992), modify the initial estimates so as to ensure consistency between the robust estimates obtained at the domain level and the robust estimate obtained at the population level. The problem of

consistency for domains has been studied in the context of small area estimation; for example, see You, Rao and Dick (2004) and Datta, Gosh, Steorts and Maple (2011).

We conclude this section with a discussion of the concept of robustness in classical statistics and robustness in finite populations. In classical statistics, we deal with infinite populations, for which we want to estimate the mean, say. In this context, an outlier is a value that was generated under a different model from the one under which the majority of the observations were generated. The presence of outliers in the sample can be attributed to the fact that the population from which the sample is generated is a mixture of distributions or that some observations are subject to measurement errors. In classical statistics, we usually want to conduct inferences about the population of inliers. The aim is therefore to construct estimators that are robust in the sense that they are not seriously affected by the presence of outliers in the sample. In this context, it is desirable to construct robust estimators that have a high breakdown point and/or a bounded influence function. In finite populations, measurement errors are corrected at the verification stage, and it is assumed that there are none left at the estimation stage. The aim is to conduct an inference about the "total" population, which includes both outliers and inliers. In other words, in contrast to classical statistics, we are not just interested in the population of inliers. In this context, estimators that have a high breakdown point and/or a bounded influence function are generally not appropriate because they can lead to large biases. We will give preference to estimators that are robust in the sense that (i) they are more stable than classical estimators in the presence of influential values and almost as efficient as classical estimators in their absence, and (ii) they converge to classical estimators as the sample size and the population size increase. Simulation studies are presented in Section 6. Section 7 concludes with a discussion.

2 Measure of influence: Conditional bias

Consider a finite population of individuals, denoted by U, of size N. We want to estimate the total for the variable of interest y, denoted by $t = \sum_{i \in U} y_i$. From the population we select a sample S, of (expected) size n, using the sampling design $p(S)$. A classical estimator of t is the expansion estimator, also known as the Horvitz-Thompson estimator, $\hat{t} = \sum_{i \in S} d_i y_i$, where $d_i = 1/\pi_i$ is the sampling weight of unit i and π_i denotes its probability of inclusion in the sample. Although the expansion estimator, $\hat{t}$, is design-unbiased for t, it can be highly unstable in the presence of influential values.

To measure the impact (or influence) that a sampled unit has on the expansion estimator, we use the concept of conditional bias of a unit; see Moreno-Rebollo, Muñoz-Reyez and Muñoz-Pichardo (1999), Moreno-Rebollo, Muñoz-Reyez, Jimenez-Gamero and Muñoz-Pichardo (2002) and Beaumont, Haziza and Ruiz-Gazen (2013). Let I_i be the sample selection indicator variable for unit i such that $I_i = 1$ if $i \in S$ and $I_i = 0$, otherwise. The conditional bias of the estimator $\hat{t}$ associated with a sampled unit is defined as

$$B_{1i}^{\mathrm{HT}} = E_p\left(\hat{t} \,|\, I_i = 1\right) - t = \sum_{j \in U} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) y_j, \tag{2.1}$$

where π_{ij} is the joint probability of inclusion of units i and j in the sample. In general, the conditional bias (2.1) is unknown, since the values of the variable of interest are observed only for the sampled units. In practice, the conditional bias must be estimated. We consider the conditionally unbiased estimator (for example, see Beaumont et al. 2013):

$$\begin{aligned}\hat{B}_{1i}^{\mathrm{HT}} &= \sum_{j\in S}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_j\pi_{ij}}\right)y_j \\ &= (d_i-1)y_i+\sum_{j\in S, j\neq i}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_j\pi_{ij}}\right)y_j.\end{aligned} \tag{2.2}$$

This estimator is conditionally unbiased in the sense that $E_p\left(\hat{B}_{1i}^{\mathrm{HT}}|I_i=1\right)=B_{1i}^{\mathrm{HT}}$. We make the following remarks on the conditional bias and its estimator: (i) The conditional bias (2.1) and its estimator (2.2) depend on the inclusion probabilities π_i and the joint inclusion probabilities π_{ij}. In other words, the conditional bias is a measure that takes the sampling design into account. (ii) If $\pi_i=1$, then $B_{1i}^{\mathrm{HT}}=0$ and, similarly, $\hat{B}_{1i}^{\mathrm{HT}}=0$. That is, when $\pi_i=1$, unit i is selected in all possible samples, and consequently $E_p\left(\hat{t}|I_i=1\right)-t=E_p\left(\hat{t}\right)-t=0$, since $\hat{t}$ is a design-unbiased estimator of t. A unit selected systematically in the sample therefore has no influence and does not contribute to the variance of $\hat{t}$. (iii) The estimated conditional bias (2.2) depends on the second-order inclusion probabilities, π_{ij}. For some designs, these probabilities may be difficult to calculate, in which case approximations will be used. For sampling designs that belong to the class of high-entropy designs (e.g., Berger 1998), a number of approximations of the second-order inclusion probabilities have been proposed in the literature; for example, see Haziza, Mecatti and Rao (2008). An alternative solution is to calculate approximations of the π_{ij} using Monte Carlo methods; see Fattorini (2006) and Thompson and Wu (2008).

For a stratified simple random sampling design, the conditional bias (2.1) associated with sampled unit i in stratum h is given by

$$B_{1i}^{\mathrm{HT}}=\frac{N_h}{N_h-1}\left(\frac{N_h}{n_h}-1\right)(y_i-\bar{y}_{Uh}), \tag{2.3}$$

where n_h denotes the size of the sample selected in stratum $h, \bar{y}_{Uh}=N_h^{-1}\sum_{i\in U_h}y_i$, and U_h denotes the population of units in stratum h of size $N_h, h=1,\ldots,H$. The estimator of the conditional bias (2.2) reduces to

$$\hat{B}_{1i}^{\mathrm{HT}}=\frac{n_h}{n_h-1}\left(\frac{N_h}{n_h}-1\right)(y_i-\bar{y}_{Sh}),$$

where $\bar{y}_{Sh}=n_h^{-1}\sum_{i\in S_h}y_i$ and S_h is the sample in stratum h.

For a Poisson design, the conditional bias of sampled unit i is given by

$$B_i^{\mathrm{HT}}(I_i=1)=(d_i-1)y_i. \tag{2.4}$$

In contrast to the simple random sampling without-replacement design, the conditional bias (2.4) is known for all units in the sample, since it does not depend on finite population parameters.

3 Robust estimation based on the conditional bias

To guard against the undue influence of certain units, it is advisable to construct robust estimators of the total t, that is, estimators that reduce the impact of the most influential units. We consider a class of estimators of the form

$$\hat{t}_R = \hat{t} + \Delta, \tag{3.1}$$

where Δ is a certain random variable. As we will see in Section 4, the winsorized estimators considered can be written in form (3.1). As in Beaumont et al. (2013), we want to determine the value of Δ that minimizes the maximum estimated conditional bias of $\hat{t}_R$ in the sample. Formally, we are seeking the value of Δ that minimizes

$$\max_{i \in S} \left\{ \left| \hat{B}_{1i}^R \right| \right\}, \tag{3.2}$$

where $\hat{B}_{1i}^R$ denotes the estimated conditional bias of $\hat{t}_R$ associated with sampled unit i. This conditional bias is given by

$$\begin{aligned} B_{1i}^R &= E_p\left(\hat{t}_R \,|I_i = 1\right) - t \\ &= B_{1i}^{\mathrm{HT}} + E_p\left(\Delta \,|I_i = 1\right) \end{aligned} \tag{3.3}$$

which is estimated by

$$\hat{B}_{1i}^R = \hat{B}_{1i}^{\mathrm{HT}} + \Delta, \tag{3.4}$$

where $\hat{B}_{1i}^{\mathrm{HT}}$ is a conditionally unbiased estimator of B_{1i}^{HT}. If we note that Δ is a conditionally unbiased estimator of $E_p\left(\Delta \,|I_i = 1\right)$, it follows that the estimator of the conditional bias (3.4) is conditionally unbiased for B_{1i}^R. In other words, we have $E_p\left\{\hat{B}_{1i}^R \,|I_i = 1\right\} = B_{1i}^R$.

Beaumont et al. (2013) showed that the value of Δ that minimizes (3.2) is given by

$$\Delta_{\mathrm{opt}} = -\frac{1}{2}\left(\hat{B}_{\min} + \hat{B}_{\max}\right),$$

where $\hat{B}_{\min} = \min_{i \in S}\left(\hat{B}_{1i}^{\mathrm{HT}}\right)$ and $\hat{B}_{\max} = \max_{i \in S}\left(\hat{B}_{1i}^{\mathrm{HT}}\right)$. Estimator (3.1) then becomes

$$\hat{t}_R = \hat{t} - \frac{1}{2}\left(\hat{B}_{\min} + \hat{B}_{\max}\right). \tag{3.5}$$

Beaumont et al. (2013) demonstrated that under certain regularity conditions, the estimator (3.5) is design-consistent; that is, $\hat{t}_R - t = O_p\left(N/\sqrt{n}\right)$.

4 Application to winsorized estimators

Estimator (3.5) can be written in alternative forms, which can make it easier to implement in some cases. We consider the winsorized form. This form has been widely studied in the literature. As mentioned in Section 1, standard winsorization is distinguished from Dalén-Tambay winsorization.

Standard winsorization involves decreasing the value of units that are above a particular threshold, taking their weight into account. Let $\tilde{y}_i$ be the value of variable y for unit i after winsorization. We have

$$\tilde{y}_i = \begin{cases} y_i & \text{if } d_i y_i \leq K \\ \dfrac{K}{d_i} & \text{if } d_i y_i > K \end{cases} \tag{4.1}$$

where $K > 0$ is the winsorization threshold. The standard winsorized estimator of the total t is given by

$$\begin{aligned} \hat{t}_s &= \sum_{i \in S} d_i \tilde{y}_i \\ &= \hat{t} + \Delta(K), \end{aligned} \tag{4.2}$$

where

$$\Delta(K) = -\sum_{i \in S} \max(0, d_i y_i - K).$$

Hence, the estimator (4.2) can be written in the form (3.1). An alternative is to express $\hat{t}_s$ as a weighted sum of the initial values using modified weights:

$$\hat{t}_s = \sum_{i \in S} \tilde{d}_i y_i,$$

where

$$\tilde{d}_i = d_i \frac{\min\left(y_i, \dfrac{K}{d_i}\right)}{y_i}. \tag{4.3}$$

If $\min(y_i, K/d_i) = y_i$ (that is, if unit i is not influential), then $\tilde{d}_i = d_i$. Thus, the weight of a non-influential unit is not modified. In contrast, the modified weight of an influential unit is less than d_i and may even be less than 1. It is worth noting that a unit with a value of $y_i = 0$ presents no particular problems, since its contribution to the estimated total, $\hat{t}_s$, is zero. In this case, an arbitrary value can be assigned to the modified weight $\tilde{d}_i$.

In the case of Dalén-Tambay winsorization, the values of the variable of interest after winsorization are defined by

$$\tilde{y}_i = \begin{cases} y_i & \text{if } d_i y_i \leq K \\ \dfrac{K}{d_i} + \dfrac{1}{d_i}\left(y_i - \dfrac{K}{d_i}\right) & \text{if } d_i y_i > K \end{cases}. \tag{4.4}$$

This leads to the winsorized estimator of the total t_y :

$$\begin{aligned}\hat{t}_{\mathrm{DT}} &= \sum_{i\in S} d_i \tilde{y}_i \\ &= \hat{t} + \Delta(K),\end{aligned} \tag{4.5}$$

where

$$\Delta(K) = -\sum_{i\in S} \frac{(d_i - 1)}{d_i} \max(0, d_i y_i - K).$$

Estimator (4.5) can also be written in the form (3.1). As in the case of $\hat{t}_s$, an alternative is to express $\hat{t}_{\mathrm{DT}}$ as a weighted sum of the initial values using modified weights:

$$\hat{t}_{\mathrm{DT}} = \sum_{i\in S} \tilde{d}_i y_i,$$

where

$$\tilde{d}_i = 1 + (d_i - 1)\frac{\min\left(y_i, \dfrac{K}{d_i}\right)}{y_i}. \tag{4.6}$$

As in the case of the standard winsorized estimator, the weight of a non-influential unit is not modified. Unlike standard winsorization, Dalén-Tambay winsorization guarantees that the modified weights will not be less than 1. Once again, a unit with a value of $y_i = 0$ presents no particular problems, since its contribution to the estimated total, $\hat{t}_{\mathrm{DT}}$, is zero. In this case, an arbitrary value can be assigned to the modified weight $\tilde{d}_i$.

Since the standard and Dalén-Tambay winsorized estimators are of the form (3.1), the optimal constant K_{opt} that minimizes (3.2) is obtained by solving

$$\Delta(K) = -\frac{1}{2}\left(\hat{B}_{\min} + \hat{B}_{\max}\right)$$

or

$$\sum_{j\in S} a_j \max\left(0, d_j y_j - K\right) = \frac{\hat{B}_{\min} + \hat{B}_{\max}}{2}, \tag{4.7}$$

where $a_j = 1$ in the case of $\hat{t}_s$ and $a_j = \left(d_j - 1\right)/d_j$ in the case of $\hat{t}_{\mathrm{DT}}$. It is shown in the Appendix that a solution to equation (4.7) exists under the following conditions:

1. $\pi_{ij} - \pi_i\pi_j \leq 0$; and
2. $\frac{1}{2}\left(\hat{B}_{\min} + \hat{B}_{\max}\right) \geq 0.$

Condition 1 is satisfied for most one-stage designs used in practice, such as stratified simple random sampling and Poisson sampling. Condition 2 implies that $\hat{t}_R$ must be less than or equal to $\hat{t}$, since by construction, a winsorized estimator cannot be greater than the Horvitz-Thompson estimator. It is generally expected that Condition 2 will be satisfied in most skewed populations encountered in business surveys and social surveys. It is also shown in the Appendix that the solution to equation (4.7) is unique if the above conditions are met and if $y_i \geq 0$ for $i \in S$. The Appendix contains a brief description of an algorithm for finding the solution to equation (4.7).

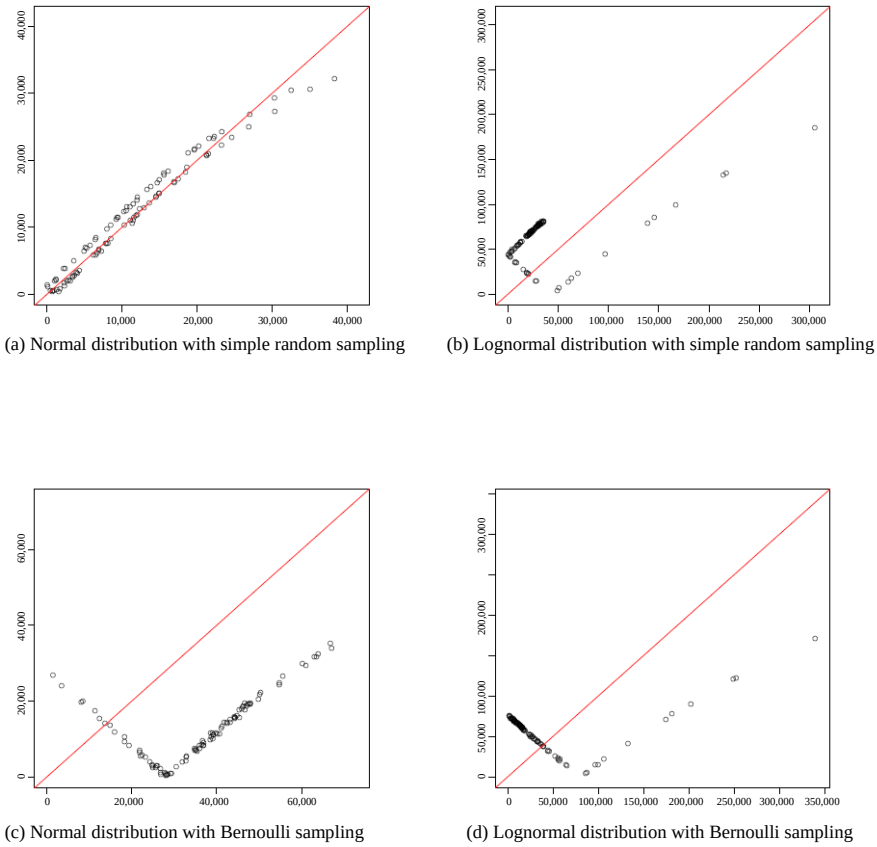
It should be noted that while the value K_{opt} is different for each type of winsorized estimator used, the resulting robust estimators are identical. In other words, we have

$$\hat{t}_s\left(K_{\text{opt}}\right) = \hat{t}_{\text{DT}}\left(K_{\text{opt}}\right) = \hat{t}_R = \hat{t} - \frac{\hat{B}_{\min} + \hat{B}_{\max}}{2}. \tag{4.8}$$

To compare the influence of each population unit with respect to the (non-robust) expansion estimator, $\hat{t}$, and its robust version (4.8), we carried out a simulation study. For that purpose, we generated two populations, each of size $N = 100$. One population was generated according to a normal distribution with mean 4,108 and standard deviation 1,500, and the other was generated according to a lognormal distribution with mean 4,108 and standard deviation 7,373. From each population we selected $M = 500{,}000$ samples according to two sampling designs: (i) a simple random sampling without-replacement design of size $n = 10$, and (ii) a Bernoulli design of expected size $n = 10$. First, we calculated the conditional bias of the Horvitz-Thompson estimator for a simple random sampling without-replacement design, given in (2.3) and for a Bernoulli design, given in (2.4). Note that the conditional bias of the Horvitz-Thompson estimator does not have to be approximated by simulation since all the population parameters are known. The conditional bias associated with unit i of the robust estimator given in (3.3) was approximated as follows: Out of the 500,000 selected samples, we identified those which contained unit i. In each of these samples, we calculated the error, $\hat{t}_R - t$. Finally, we calculated the average value of $\hat{t}_R - t$ over all the samples containing unit i.

The results for the simple random sampling without-replacement design for the normal and lognormal distributions are shown in Figures 4.1 (a) and 4.1 (b) respectively. The results for the Bernoulli sampling design for the normal and lognormal distributions are shown in Figures 4.1 (c) and 4.1 (d) respectively. In each figure, the absolute value of the conditional bias of $\hat{t}_R$ is shown in relation to the absolute value of the conditional bias of $\hat{t}$ for each population unit. The units above the first bisectrix have a conditional bias associated with $\hat{t}_R$ whose absolute value is greater than that of the conditional bias associated with $\hat{t}$. Looking first at the results for simple random sampling without replacement, we see that the behaviour of the absolute value of the conditional bias of $\hat{t}_R$ is similar to that of the absolute value of the conditional bias of $\hat{t}$, which indicates that the influence of the units is not altered significantly after robustification of the expansion estimator. This result is not surprising since the population does not contain any highly influential units. In the case of the lognormal distribution, we see that the influence of the values that have a high conditional bias associated with $\hat{t}$ has been reduced significantly. On the other hand, we note that for the majority of the data, the conditional bias of $\hat{t}_R$ is slightly higher than that of $\hat{t}$. Turning to the results for Bernoulli sampling, we see that in the case of the normal population, the influence of most units has been reduced, since the absolute value of the conditional bias of $\hat{t}_R$ is significantly lower than the

absolute value of the conditional bias of $\hat{t}$. In the case of the lognormal distribution, the results are similar to those obtained with simple random sampling without replacement for the same distribution.



(a) Normal distribution with simple random sampling

(b) Lognormal distribution with simple random sampling

(c) Normal distribution with Bernoulli sampling

(d) Lognormal distribution with Bernoulli sampling

Figure 4.1 Absolute value of the conditional biases of the robust and non-robust estimators

5 Robust estimation of domain totals

In practice, we usually want to produce estimates for population domains as well as an estimate at the global level. Let $t_g = \sum_{i \in U_g} y_i$ be the total of the $y-$variable in domain g. We assume that the domains form a partition of the population such that $t = \sum_{i \in U} y_i = \sum_{g=1}^{G} t_g$, where G is the number of domains. Let S_g be the set of sampled units in domain g. The expansion estimator of t_g is given by $\hat{t}_g = \sum_{i \in S_g} d_i y_i$. We have the consistency relation $\sum_{g=1}^{G} \hat{t}_g = \hat{t}$.

In the presence of influential values, we can apply a robust procedure separately for each domain using the method described in Section 3, which leads to G robust estimators, $\hat{t}_{R,g}$. A robust estimator of the

total at the population level, $\hat{t}_{R(\text{agg})}$, is easily obtained by aggregating the robust estimators $\hat{t}_{R,g}$. Thus, we have $\hat{t}_{R(\text{agg})} = \sum_{g=1}^{G} \hat{t}_{R,g}$. The consistency relation between the domain-level estimates and the population-level estimate is therefore satisfied. However, aggregating G robust estimators, each suffering from a potential bias, may produce a highly biased aggregate robust estimator, $\hat{t}_{R(\text{agg})}$. In most cases, the bias of $\hat{t}_{R(\text{agg})}$ will be negative, since each of the $\hat{t}_{R,g}$ estimators has a negative bias.

To avoid having an estimator with an unacceptable bias, we first compute the robust estimator (4.8), $\hat{t}_{R,g}$, for each domain. Then, we independently compute a robust estimator of the total t in the population, $\hat{t}_{R,0}$, given by (4.8). In this case, however, the consistency relation is no longer necessarily satisfied. In other words, we have $\hat{t}_{R,0} \neq \sum_{g=1}^{G} \hat{t}_{R,g}$, in general. It is therefore necessary to force consistency between the robust domain estimates and the aggregate robust estimate using a method similar to calibration. To do so, we compute final robust estimates $\hat{t}_{R,g}^{*}, g = 0,1,..,G,$ that are as close as possible to the initial robust estimates $\hat{t}_{R,g}$, based on a particular distance function, and that satisfy the calibration equation

$$\sum_{g=1}^{G} \hat{t}_{R,g}^{*} = \hat{t}_{R,0}^{*}. \tag{5.1}$$

In the case of the generalized chi-square distance function, we are seeking final robust estimates, $\hat{t}_{R,g}^{*}$, such that

$$\sum_{g=0}^{G} \frac{\left\{\hat{t}_{R,g}^{*} - \hat{t}_{R,g}\right\}^2}{2q_g \hat{t}_{R,g}} \tag{5.2}$$

is minimized subject to (5.1). The coefficient q_g in the above expression is a weight assigned to the initial estimate in domain $g, \hat{t}_{R,g}$, and is interpreted as its importance in the minimization problem. Using the Lagrange multipliers method, we can easily obtain a solution to this minimization problem. The solution is given by

$$\hat{t}_{R,g}^{*} = \hat{t}_{R,g} - \frac{\sum_{h=0}^{G} \delta_h \hat{t}_{R,h}}{\sum_{h=0}^{G} q_h \hat{t}_{R,h}} \delta_g q_g \hat{t}_{R,g}, \tag{5.3}$$

where $\delta_0 = -1$ and $\delta_g = 1$, for $g = 1,\ldots,G$.

We make the following remarks: (i) If $q_g = 0$, then the final robust estimate $\hat{t}_{R,g}^{*}$ is identical to the initial robust estimate $\hat{t}_{R,g}$. Thus, if we want to ensure that the initial estimate in domain g is not modified excessively, we simply associate it with a small value of q_g. This point is also illustrated empirically in Section 6.2. (ii) Note that like the initial robust estimates at the domain level, $\hat{t}_{R,g}$, for $g = 1,\ldots,G$, the initial robust estimate at the population level, $\hat{t}_{R,0}$, can also be modified. (iii) If $q_0 = 0$

(in other words, the initial robust estimate for the population level is not modified) and $q_g = q$ for $g = 1, \ldots, G,$ where q is a strictly positive constant, expression (5.3) simplifies to

$$\hat{t}^*_{R,g} = \hat{t}_{R,g}\left(\frac{\hat{t}_{R,0}}{\hat{t}_{R(\text{agg})}}\right). \tag{5.4}$$

In this case, the initial estimates $\hat{t}_{R,g}$ are all modified by the same factor, $\hat{t}_{R,0}/\hat{t}_{R(\text{agg})}$. (iv) How can we set the values of q_g in practice? It seems natural to adopt the following choice:

$$q_g = \widehat{\text{CV}}\left(\hat{t}_g\right)\Big/\sum_{g=1}^{G}\widehat{\text{CV}}\left(\hat{t}_g\right),$$

where $\widehat{\text{CV}}\left(\hat{t}_g\right)$ is the estimated coefficient of variation (CV) associated with domain g. For example, in a repeated survey, the estimated CV observed in a previous iteration can be used. This choice of q_g is based on the fact that we will not want to make a large change in the initial estimate associated with a domain that has a small estimated CV. In such a domain, the problem of influential values is clearly less serious, and the initial robust estimate $\hat{t}_{R,g}$ is expected to be relatively close to the actual total t_g. In other words, the robust estimator $\hat{t}_{R,g}$ should have low bias and be relatively stable. It therefore makes sense not to attempt to change the initial robust estimate substantially. (v) In (5.2), we used the generalized chi-square distance, which leads to the linear method. In the literature on calibration (e.g., Deville and Särndal 1992), there are a number of other calibration methods. In particular, there is the Kullback-Leibler distance, which leads to the exponential method and the logit and truncated linear methods. Using the last two methods, we can specify positive bounds C_1 and C_2 such that $C_1 \leq \hat{t}^*_{R,g}/\hat{t}_{R,g} \leq C_2$. In other words, we ensure that the ratio $\hat{t}^*_{R,g}/\hat{t}_{R,g}$ falls within the interval between C_1 and C_2. Note that the calibration procedure may lead to $\hat{t}^*_{R,g} - \hat{t}_g \geq 0,$ for a certain $g,$ which is counterintuitive. In this case, we simply include the constraint $\hat{t}^*_{R,g} \leq \hat{t}_g$ for $g = 1, \ldots, G,$ in the calibration procedure. (vi) An alternative is to express $\hat{t}^*_{R,g}$ as a weighted sum of the initial values using modified weights:

$$\hat{t}^*_{R,g} = \sum_{i \in S_g} \tilde{d}_i{}^* y_i,$$

where

$$\tilde{d}_i{}^* = \tilde{d}_i\left(1 - \delta_g q_g \frac{\sum_{h=0}^{G}\delta_h \hat{t}_{R,h}}{\sum_{h=0}^{G} q_h \hat{t}_{R,h}}\right)$$

and $\tilde{d}_i$ is given by either (4.3) or (4.6). We can also write the estimator $\hat{t}^*_{R,g}$ as a weighted sum with the initial weights using modified values:

$$\hat{t}^*_{R,g} = \sum_{i \in S_g} d_i \tilde{y}^*_i,$$

where

$$\tilde{y}_i^* = \tilde{y}_i \left(1 - \delta_g q_g \frac{\sum_{h=0}^{G} \delta_h \hat{t}_{R,h}}{\sum_{h=0}^{G} q_h \hat{t}_{R,h}} \right), \quad i \in g$$

and $\tilde{y}_i$ is given by either (4.1) or (4.4). (vii) We may want to find the winsorization thresholds $K_g, g = 1, \ldots, G,$ such that the standard winsorized estimator or the Dalén-Tambay winsorized estimator is equal to $\hat{t}_{R,g}^*$. We can follow a procedure similar to the one in Section 4, and we can use an algorithm similar to the one in the Appendix. A necessary condition for the existence of a solution is that $\hat{t}_g - \hat{t}_{R,g}^* \geq 0$. (viii) With the proposed calibration procedure, more than one partition of the population can be dealt with jointly. For example, we may be interested in publishing both provincial estimates and industry estimates. If so, we simply insert the following calibration equations into the calibration procedure:

$$\sum_{g=1}^{G} \hat{t}_{R,g}^* = \hat{t}_{R,0}^*,$$

$$\sum_{l=1}^{L} \hat{t}_{R,l}^* = \hat{t}_{R,0}^*,$$

where G and L denote the number of provinces and the number of industries respectively. The method can also be applied to more than two partitions of the population.

6 Simulation studies

6.1 Winsorization in a simple random sampling without-replacement design

We carried out a simulation study to examine the properties of several robust estimators using 11 populations. The first 10 of size $N = 5{,}000$ consists of a variable of interest y. In each population, the $y-$values were generated according to the following model:

$$Y_i = U_i + \delta_i V_i,$$

where U_i, δ_i and V_i are random variables whose distributions are described in Table 6.1. Population 1 was generated according to a normal distribution. Populations 2 through 5 were generated using a mixture of normal distributions with contamination rates ranging from 0.5% to 5%. Populations 6 through 8 were generated according to skewed distributions. Populations 9 and 10 were generated using a mixture of lognormal distributions with contamination rates equal to 0.5% and 5%. Population 11 of size $N = 5{,}000$ is from the information technology survey produced by the French National Institute for Statistics and Economic Studies (INSEE) in 2011. One of the survey's objectives is to estimate the e-commerce sales of French companies. We use the "sales" variable in our simulation. The distribution of y in each

population is plotted in Figure 6.1. In addition, Table 6.2 presents a number of descriptive statistics for each of the populations used. For confidentiality reasons, the units for Population 11 are not shown in the plot. Similarly, there are no descriptive statistics for Population 11 in Table 6.2.

In each population, we selected $M = 5{,}000$ samples according to a simple random sampling without-replacement design of size $n = 100, 300$ and 500. For each sample, we calculated the expansion estimator $\hat{t}$ and the robust estimator (4.8). Let $y_{(1)}, \ldots, y_{(n)}$ be the values of the $y-$variable arranged in ascending order. We also calculated the first-, second- and third-order winsorized estimators, where the $p^{\text{th}}-$order winsorized estimator is obtained by replacing the p largest values in the sample with the value $y_{(n-p)}$, $p = 1, 2, 3$. In a classical statistical context, Rivest (1994) showed that the first-order winsorized estimator has good mean-square-error properties for a large class of skewed distributions.

As a measure of the bias of an estimator $\hat{\theta}$, we calculated the Monte Carlo relative bias (in percentage):

$$\text{BR}_{\text{MC}}\left(\hat{\theta}\right) = \frac{\frac{1}{M}\sum_{m=1}^{M}\left(\hat{\theta}_{(m)} - t\right)}{t} \times 100,$$

where $\hat{\theta}_{(m)}$ denotes the estimator $\hat{\theta}$ in sample $m, m = 1, \ldots, 5{,}000$. We also calculated the relative efficiency of the robust estimators with respect to the expansion estimator, $\hat{t}$:

$$\text{RE}_{\text{MC}}\left(\hat{\theta}\right) = \frac{\frac{1}{M}\sum_{m=1}^{M}\left(\hat{\theta}_{(m)} - t\right)^2}{\frac{1}{M}\sum_{m=1}^{M}\left(\hat{t}_{(m)} - t\right)^2} \times 100.$$

The results are shown in Table 6.3.

The results presented in Table 6.3 show that the once-winsorized estimator has lower bias and is generally more efficient than the two times and three times winsorized estimators, which is consistent with the results obtained by Rivest (1994). It is interesting to compare the robust estimator $\hat{t}_R$ and the once-winsorized estimator. In the case of Population 1, which does not contain any influential values, we see that both estimators have low bias and are as efficient as the expansion estimator. In the case of the populations with a mixture of normal distributions (Populations 2 to 5), we observe that the once-winsorized estimator is less efficient than the robust estimator in every scenario except for Population 5 with $n = 300$. In fact, the once-winsorized estimator is less efficient than the expansion estimator in every scenario except for Population 2 with $n = 100$. The robust estimator is more efficient than the expansion estimator except in Populations 4 and 5, for which we observe values of relative efficiency ranging from 91% to 102%. In the case of the populations with a mixture of lognormal distributions (Populations 9 and 10), we see that the bias and efficiency performance of the once-winsorized estimator and the robust estimator is very similar in all scenarios. The same is true for the skewed populations (Populations 6 to 8), for which the two estimators produce similar results. In the case of Population 11, the robust estimator has a lower bias than the once-winsorized estimator for $n = 100$, though it is less

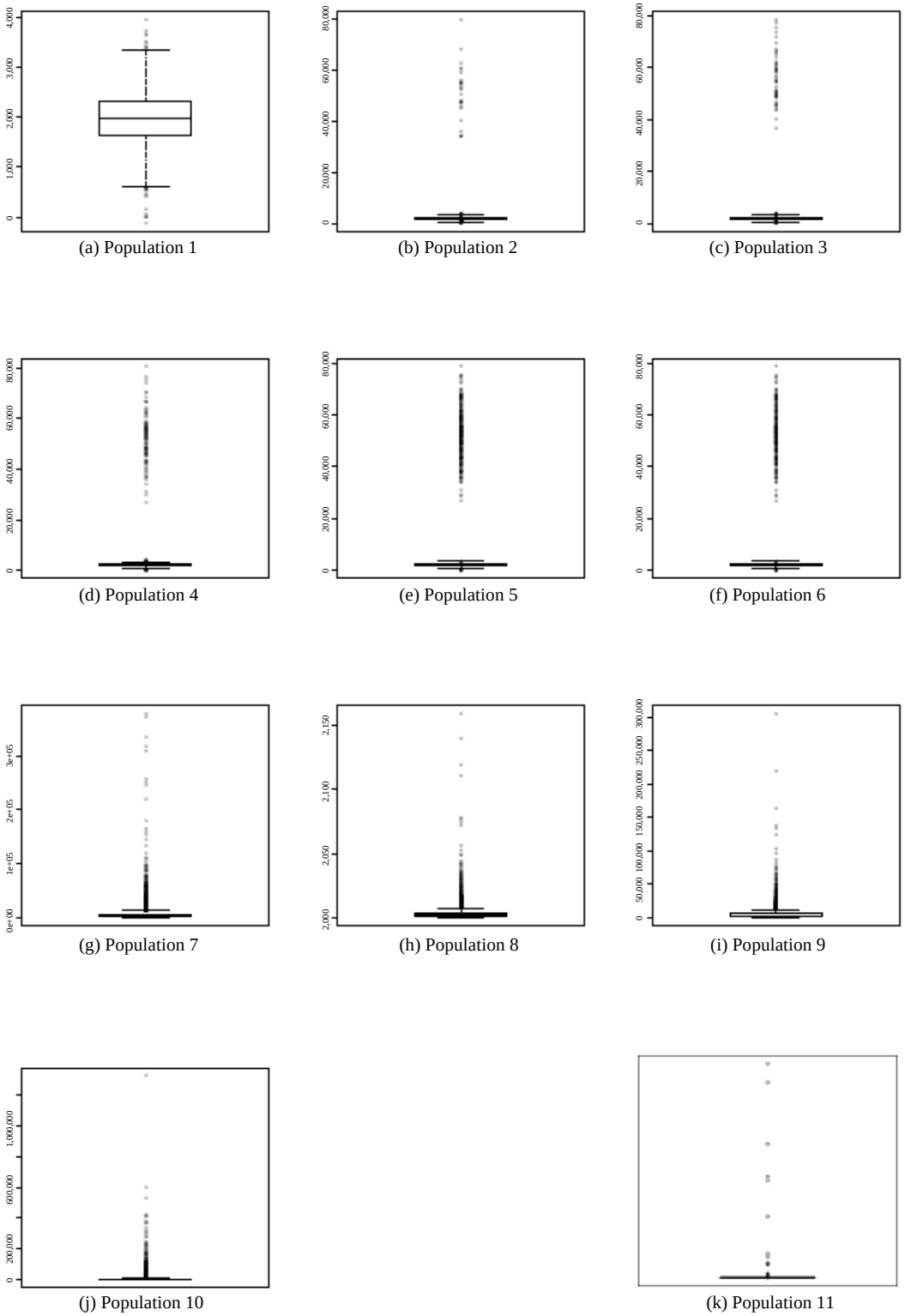
efficient (41% versus 47%). For $n = 300$ and $n = 500$, the robust estimator has a lower bias and is significantly more efficient than the once-winsorized estimator.

Table 6.1
Models used to generate the populations

Population	U_i distribution	Mixture	δ_i distribution	V_i distribution
1	$\mathcal{N}(2{,}000;\ 500)$	No		
2	$\mathcal{N}(2{,}000;\ 500)$	Yes	$\mathcal{B}(0.005)$	$\mathcal{N}(50{,}000;\ 10{,}000)$
3	$\mathcal{N}(2{,}000;\ 500)$	Yes	$\mathcal{B}(0.01)$	$\mathcal{N}(50{,}000;\ 10{,}000)$
4	$\mathcal{N}(2{,}000;\ 500)$	Yes	$\mathcal{B}(0.02)$	$\mathcal{N}(50{,}000;\ 10{,}000)$
5	$\mathcal{N}(2{,}000;\ 500)$	Yes	$\mathcal{B}(0.05)$	$\mathcal{N}(50{,}000;\ 10{,}000)$
6	$\mathcal{L}\text{og} - \mathcal{N}(\log(2{,}000);1.2)$	No		
7	$\mathcal{L}\text{og}-\mathcal{N}(\log(2{,}000);1.5)$	No		
8	$\mathcal{F}\text{rechet}(2{,}000;2.5;2.1)$	No		
9	$\mathcal{L}\text{og}-\mathcal{N}(\log(2{,}000);1.2)$	Yes	$\mathcal{B}(0.05)$	$\mathcal{L}\text{og}-\mathcal{N}(\log(5{,}000);\ 1.2)$
10	$\mathcal{L}\text{og}-\mathcal{N}(\log(2{,}000);1.2)$	Yes	$\mathcal{B}(0.05)$	$\mathcal{L}\text{og}-\mathcal{N}(\log(5{,}000);\ 1.2)$

Table 6.2
Descriptive statistics for the ten simulated populations

Descriptive	Population									
statistic	1	2	3	4	5	6	7	8	9	10
min	132.3	314.9	105.3	275.9	187.4	23.6	7.6	2,000.9	20.5	26.6
max	3,968	79,506	78,526	80,540	78,690	252,612	379,751	2,159	305,612	1.3×10^{6}
$Q1$	1,639	1,667	1,664	1,666	1,685	883	743	200	920	913
Median	1,986	1,993	1,997	2,015	2,053	1,996	1,981	2,002	2,167	2,041
$Q3$	2,330	2,337	2,339	2,349	2,421	4,505	5,337	2,004	5,018	4,927
Mean	1,985	2,267	2,536	2,976	4,661	4,005	6,118	2,004	4,738	7,883
Standard deviation	503	3,709	5,506	7,119	11,470	7,353	17,190	5.89	9,796	33,111
Skewness	0.0	14.0	10.2	7.3	4.3	4.2	11.6	11.8	12.1	18.4
Kurtosis	3	209	109	56	20	19	196	228	267	570
CV	0.25	1.6	2.2	2.4	2.5	1.8	2.8	2.9×10^{-3}	2.0	4.2



(a) Population 1 (b) Population 2 (c) Population 3

(d) Population 4 (e) Population 5 (f) Population 6

(g) Population 7 (h) Population 8 (i) Population 9

(j) Population 10 (k) Population 11

Figure 6.1 Distribution of the variable of interest in the 11 populations.

Table 6.3
Monte Carlo relative bias (in %) and relative efficiency (in parentheses) of several estimators

Population	n	$\hat{t}_R$	Winsorization		
			Once	Two times	Three times
1	100	-0.1(100)	-0.1(100)	-0.2(101)	-0.3(102)
	300	0.0(100)	-0.0(100)	-0.0(100)	-0.1(100)
	500	0.0(100)	-0.0(100)	-0.0(100)	-0.0(100)
2	100	-4.9(59)	-7.5(87)	-10.7(65)	-11.9(55)
	300	-2.9(87)	-3.0(129)	-6.8(158)	-9.5(169)
	500	-1.9(96)	-1.2(122)	-3.6(175)	-6.5(226)
3	100	-6.9(74)	-8.9(122)	-16.5(119)	-20.0(107)
	300	-3.5(99)	-1.9(122)	-5.6(171)	-10.6(232)
	500	-2.4(102)	-0.9(107)	-2.2(130)	-4.5(186)
4	100	-7.6(91)	-6.2(131)	-15.5(169)	-24.4(194)
	300	-2.9(101)	-0.6(103)	-2.1(118)	-4.4(154)
	500	-2.0(102)	-0.6(102)	-1.1(101)	-1.8(108)
5	100	-5.7(102)	-1.1(104)	-4.1(126)	-9.7(173)
	300	-2.2(102)	-0.4(100)	-0.8(101)	-1.4(102)
	500	-1.2(100)	-0.1(100)	-0.3(100)	-0.5(101)
6	100	-5.7(79)	-5.4(75)	-8.2(80)	-10.6(89)
	300	-2.6(84)	-2.6(79)	-3.9(81)	-5.1(88)
	500	-2.0(86)	-2.0(81)	-3.0(82)	-3.8(88)
7	100	-8.4(72)	-9.3(73)	-14.7(72)	-18.7(79)
	300	-4.5(86)	-4.4(95)	-7.8(91)	-10.2(95)
	500	-3.5(94)	-3.1(105)	-6.0(106)	-8.1(109)
8	100	-0.0(69)	-0.0(75)	-0.0(77)	-0.0(85)
	300	-0.0(82)	-0.0(88)	-0.0(87)	-0.0(95)
	500	-0.0(88)	-0.0(96)	-0.0(94)	-0.0(100)
9	100	-5.7(73)	-5.8(71)	-9.5(72)	-12.4(80)
	300	-3.5(87)	-3.5(85)	-5.4(88)	-6.8(98)
	500	-2.4(88)	-2.4(88)	-3.8(90)	-4.9(97)
10	100	-13.5(68)	-15.0(70)	-24.6(76)	-31.7(89)
	300	-7.5(80)	-7.2(79)	-12.1(85)	-16.3(97)
	500	-5.3(85)	-5.1(83)	-8.4(91)	-11.4(103)
11	100	-22.8(47)	-32.6(41)	-42.0(42)	-47.7(47)
	300	-14.7(65)	-20.0(77)	-29.6(68)	-34.3(75)
	500	-11.3(76)	-14.6(96)	-24.3(90)	-29.3(97)

6.2 Winsorization in a stratified simple random sampling without-replacement design

We also tested the calibration method described in Section 5. We generated a population of size $N = 5{,}000$, which we divided into five strata, $U_1, \ldots, U_5$, of size $N_1, \ldots, N_5$, respectively; see Table 6.4 for the values of N_h. In each stratum, we generated a variable of interest y according to a lognormal distribution with parameters $\log(2{,}000)$ and 1.5.

From the population we selected $M = 5{,}000$ samples according to a stratified simple random sampling without-replacement design. In stratum U_h, we selected a sample S_h of size n_h according to a simple random sampling without-replacement design; see Table 6.4 for the sizes n_h and the corresponding sampling fractions, $f_h = n_h / N_h$.

The objective here is to estimate the total in the population, $t = \sum_{i \in U} y_i$, and the stratum totals $t_h = \sum_{i \in U_h} y_i$, $h = 1, \ldots, H$. In other words, in our example, the strata correspond to domains of interest. Since the strata form a partition of the population, we have the consistency relation, $t = \sum_{h=1}^{H} t_h$. Similarly, the expansion estimators satisfy the consistency relation $\hat{t} = \sum_{h=1}^{H} \hat{t}_h$, where $\hat{t} = \sum_{i \in S} d_i y_i$ and $\hat{t}_h = \sum_{i \in S_h} d_i y_i$ with $d_i = N_h / n_h$ if $i \in U_h$.

For each sample, we first computed the robust estimator (4.8) in each stratum and aggregated the robust estimates to produce an aggregate robust estimate, $\hat{t}_{R(\text{agg})} = \sum_{h=1}^{H} \hat{t}_{R,h}$. Independently, we computed the robust estimator (4.8), denoted $\hat{t}_{R,0}$, at the population level. To ensure that the consistency relation (5.1) was satisfied, we performed the calibration procedure described in Section 5 to obtain the final robust estimates $\hat{t}^*_{R,h}, h = 0, \ldots, 5$. We used four systems of coefficients q_h : (1) $q_0 = 0$ and $q_1 = \ldots = q_5 = 1$; (2) $q_0 = 0$ and $q_h = n_h^{-1}(1 - f_h)$, $h = 1, \ldots, 5$; (3) $q_0 = 0$ and $q_h = \text{CV}(\hat{t}_h) = \sqrt{N_h^2 (1 - f_h) n_h^{-1} S_h^2} / t_h$, where $S_h^2 = (N_h - 1)^{-1} \sum_{i \in U_h} (y_i - \bar{y}_{Uh})^2$, $h = 1, \ldots, 5$; (4) $q_0 = 0$ and $q_h = \widehat{\text{CV}}(\hat{t}_h) = \sqrt{N_h^2 (1 - f_h) n_h^{-1} s_h^2} / \hat{t}_h$, where $s_h^2 = (n_h - 1)^{-1} \sum_{i \in S_h} (y_i - \bar{y}_{Sh})^2$, $h = 1, \ldots, 5$. We make the following remarks on the choice of the coefficients q_h : (i) For all four systems, we assigned a weight $q_0 = 0$ to estimate $\hat{t}_{R,0}$, which is equivalent to making no change in the robust estimate at the population level. In other words, we have $\hat{t}^*_{R,0} = \hat{t}_{R,0}$. (ii) The first weighting system assigns an equal weight to all strata regardless of the sample size or sampling fraction. (iii) In the case of the second system, the coefficient q_h is a function of the sample size n_h and the sampling fraction f_h, but it is independent of the intra-stratum variability S_h^2. (iv) In the third and fourth systems, the choice of q_h depends on the actual CV and the estimated CV respectively, for the reasons mentioned in Section 5.

Table 6.4
Characteristics of the strata

Stratum	1	2	3	4	5
N_h	2,000	1,500	1,000	400	100
n_h	20	75	100	80	80
f_h	0.01	0.05	0.1	0.2	0.8

For each robust estimator, we computed the Monte Carlo relative bias (as a percentage) and the relative efficiency (with respect to the expansion estimator); see Section 6.1. The results are presented in Table 6.5.

The results show that the initial robust estimators $\hat{t}_{R,h}$ are biased, as expected. The bias is larger in strata with a small sampling fraction. For example, in Stratum 1, for which $f_1 = 1\%$, the relative bias of $\hat{t}_{1,h}$ is -11.9%, compared with only -1.5% in Stratum 5, for which $f_5 = 80\%$. We also note that the initial robust estimators are all more efficient than the corresponding expansion estimator, with relative

efficiency values ranging from 57% to 97%. The aggregate estimator $\hat{t}_{R(\text{agg})}$ obtained by summing the initial estimators $\hat{t}_{R,h}, h = 1,\ldots,5$ shows a modest bias with a value equal to -5.7% but is more efficient than the population-level expansion estimator $\hat{t}$, with a relative efficiency of 87%.

The population-level winsorized estimator, $\hat{t}_{R,0}$, shows a small bias with a value equal to -2.8% and is significantly more efficient than the expansion estimator, with a relative efficiency of 81%. The final estimators $\hat{t}^*_{R,h}$ obtained using the system of coefficients $q_h = 1$ for $h = 1,\ldots,5$ all have lower bias than the initial estimator $\hat{t}_{R,h}$, except for Stratum 5. This is due to the fact that we force the sum of the final estimates $\hat{t}^*_{R,h}$ to calibrate on a low-bias estimator. On the other hand, the decrease in the bias is accompanied by a slight decrease in efficiency. For example, in Stratum 4, the relative efficiency is 63% for the robust estimator $\hat{t}_{R,4}$ and 66% for the final estimator $\hat{t}^*_{R,4}$. In the case of Stratum 5, the first system of coefficients is clearly unsuitable, since it leads to a change in the estimate for this stratum, like all the other strata, when this stratum has a very high sampling fraction of 80%. In fact, for this system of coefficients, the estimator $\hat{t}^*_{R,5}$ is less efficient than the expansion estimator, with a relative efficiency of 104. The second choice of coefficients q_h, which takes the sampling fraction f_h and the sample size n_h into account, leads to some interesting results. The final robust estimator in Stratum 1, $\hat{t}^*_{R,1}$, has an appreciably lower bias than the initial estimator $\hat{t}_{R,1}$ and the final estimator based on the first system of coefficients, at the cost of a slight loss of efficiency. For Stratum 5, the estimator $\hat{t}^*_{R,5}$ has a low bias (a relative bias of -0.8%) and the same 97% efficiency as the initial estimator $\hat{t}_{R,5}$. The third and fourth q_h weighting systems lead to similar relative bias and relative efficiency results. For Stratum 1, they lead to lower relative biases than the first weighting system, at the cost of a slight loss of efficiency. For Strata 2, 3 and 4, all four systems of coefficients exhibit similar relative bias and relative efficiency. For Stratum 5, the final estimators are virtually unbiased and no less efficient that the expansion estimator.

Table 6.5
Monte Carol relative bias (in %) and relative efficiency (in parentheses) of the robust estimators at the global level and the stratum level

Global estimator		$\hat{t}_{R(\text{agg})}$	$\hat{t}_{R,0} = \hat{t}^*_{R,0}$	$\hat{t}_{R,0} = \hat{t}^*_{R,0}$	$\hat{t}_{R,0} = \hat{t}^*_{R,0}$	$\hat{t}_{R,0} = \hat{t}^*_{R,0}$
		-5.7(87)	-2.8(81)	-2.8(81)	-2.8(81)	-2.8(81)
			$\hat{t}^*_{R,h}$			
		$\hat{t}_{R,h}$	$q_h = 1$	$q_h = n_h^{-1}(1 - f_h)$	$q_h = \text{CV}(\hat{t}_h)$	$q_h = \widehat{\text{CV}}(\hat{t}_h)$
Stratum	1	-11.9(57)	-9.1(60)	-0.9(67)	-5.7(62)	-6.7(64)
	2	-6.3(74)	-3.4(76)	-3.3(76)	-3.3(76)	-3.1(78)
	3	-6.0(69)	-3.1(70)	-3.8(69)	-3.2(70)	-3.2(70)
	4	-6.6(63)	-3.7(66)	-4.2(65)	-3.3(66)	-3.4(70)
	5	-1.5(97)	1.5(104)	-0.8(97)	-0.2(98)	0.1(99)

7 Discussion

This paper outlined a proposed method for determining the threshold for winsorized estimators. This method has the advantage of being simple to apply in practice and can be used for sampling designs with

unequal probabilities. We also proposed a calibration method that satisfies a consistency relation between the domain-level winsorized estimates and a population-level winsorized estimate. Although we applied the method in the case of winsorized estimators, it can be used with any type of robust estimator.

Acknowledgements

The authors are grateful to an associate editor and two reviewers for their comments and suggestions, which substantially improved the quality of this paper. David Haziza's research was funded by a grant from the Natural Sciences and Engineering Research Council of Canada.

Appendix

We want to show that there exists a solution to the equation

$$-\Delta(K) = \sum_{j \in S} a_j \max\left(0, d_j y_j - K\right) = \frac{\hat{B}_{\min} + \hat{B}_{\max}}{2} = \hat{t} - \hat{t}_R$$

under the conditions $\pi_{ij} - \pi_i \pi_j \leq 0$ and $\frac{1}{2}\left(\hat{B}_{\min} + \hat{B}_{\max}\right) \geq 0$.

First, we arrange the units in order from the smallest value of $b_i = d_i y_i, i \in S,$ to the largest, so that unit 1 has the smallest value of b_i and unit n the largest value. We begin by considering the case of $\frac{1}{2}\left(\hat{B}_{\min} + \hat{B}_{\max}\right) = 0$. We have to solve the equation $-\Delta(K) = 0,$ and we can easily see that this equation is satisfied for all $K \geq b_n$.

We now turn to the case of $\frac{1}{2}\left(\hat{B}_{\min} + \hat{B}_{\max}\right) > 0$. We note first that the function $-\Delta(K)$ is continuous and piecewise linear for $0 \leq K \leq b_n$. The pieces are defined by the intervals $\left[b_{j-1}, b_j\right[, j = 1, \ldots, n,$ where $b_0 = 0$. We also note that $-\Delta(0) = \sum_{j=m}^{n} a_j b_j > 0,$ where m is the smallest index such that $b_m \geq 0$. By the intermediate value theorem, there is a solution to equation (4.7) if we can show that

$$-\Delta(b_n) = 0 < \frac{1}{2}\left(\hat{B}_{\min} + \hat{B}_{\max}\right) \leq -\Delta(0) = \sum_{j=m}^{n} a_j b_j. \qquad \text{(A.1)}$$

The first inequality follows directly from the condition $\frac{1}{2}\left(\hat{B}_{\min} + \hat{B}_{\max}\right) > 0$. To prove the second inequality, we first note that $\frac{1}{2}\left(\hat{B}_{\min} + \hat{B}_{\max}\right) \leq \hat{B}_{\max}$. If we use the estimator of the conditional bias (2.2) and the condition $\pi_{ij} - \pi_i \pi_j \leq 0,$ we observe that $\hat{B}_{\max} \leq (d_k - 1) y_k,$ index k being associated with the unit that has the largest estimated conditional bias. For the Dalén-Tambay winsorized estimator, the last inequality can be rewritten as $\hat{B}_{\max} \leq a_k b_k$. It follows that $a_k b_k \leq -\Delta(0) = \sum_{j=m}^{n} a_j b_j,$ which completes the proof that there is a solution to equation (4.7). For the standard winsorized estimator, we can also easily show that $\hat{B}_{\max} \leq a_k b_k$ and therefore that a solution exists. In addition, if the $y_i, i \in S,$ are all positive, the function $-\Delta(K)$ is monotonically decreasing for $0 \leq K \leq b_n$ and the solution is unique.

To find the solution K_{opt}, we find the largest index l such that $-\Delta(b_l) \geq \frac{1}{2}\left(\hat{B}_{\min} + \hat{B}_{\max}\right)$, for $l \leq n$. The solution can then be calculated by linear interpolation between points b_l and b_{l+1}; that is,

$$K_{\text{opt}} = b_l \frac{\Delta(b_{l+1}) - \Delta\left(K_{\text{opt}}\right)}{\Delta(b_{l+1}) - \Delta(b_l)} + b_{l+1} \frac{\Delta\left(K_{\text{opt}}\right) - \Delta(b_l)}{\Delta(b_{l+1}) - \Delta(b_l)},$$

where $\Delta\left(K_{\text{opt}}\right) = -\frac{1}{2}\left(\hat{B}_{\min} + \hat{B}_{\max}\right)$.

References

Beaumont, J.-F., Haziza, D. and Ruiz-Gazen, A. (2013). A unified approach to robust estimation in finite population sampling. *Biometrika,* 100, 555-569.

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

Clark, R.G. (1995). Winsorization methods in sample surveys. Masters Thesis, Department of Statistics, Australian National University.

Dalén, J. (1987). Practical estimators of a population total which reduce the impact of large observations. R and D Report. Statistics Sweden.

Datta, G.S., Gosh, M., Steorts, R. and Maple, J. (2011). Bayesian benchmarking with applications to small area estimation. *Test,* 20, 574-588.

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

Fattorini, L. (2006). Applying the Horvitz-Thompson criterion in complex designs: A computer-intensive perspective for estimating inclusion probabilities. *Biometrika,* 93, 269-278.

Haziza, D., Mecatti, F. and Rao, J.N.K. (2008). Evaluation of some approximate variance estimators under the Rao-Sampford unequal probability sampling design. *Metron,* 66, 91-108.

Kokic, P.N., and Bell, P.A. (1994). Optimal Winsorizing cutoffs for a stratified finite population estimator. *Journal of Official Statistics,* 10, 419-435.

Moreno-Rebollo, J.L., Muñoz-Reyez, A.M. and Muñoz-Pichardo, J.M. (1999). Influence diagnostics in survey sampling: Conditional bias. *Biometrika,* 86, 923-928.

Moreno-Rebollo, J.L., Muñoz-Reyez, A.M., Jimenez-Gamero, M.D. and Muñoz-Pichardo, J. (2002). Influence diagnostics in survey sampling: Estimating the conditional bias. *Metrika,* 55, 209-214.

Rivest, L.-P. (1994). Statistical properties of Winsorized means for skewed distributions. *Biometrika,* 81, 373-383.

Rivest, L.-P., and Hidiroglou, M. (2004). Outlier treatment for disaggregated estimates. *Proceedings of the Survey Research Methods Section,* American Statistical Association, Alexandria, Virginia, 4248-4256.

Rivest, L.-P., and Hurtubise, D. (1995). On Searls' Winsorized mean for skewed populations. *Survey Methodology*, 21, 2, 107-116.

Tambay, J.-L. (1988). An integrated approach for the treatment of outliers in sub-annual surveys. *Proceedings of the Survey Research Methods Section,* American Statistical Association, Alexandria, Virginie, 229-234.

Thompson, M.E., and Wu, C. (2008). Simulation-based randomized systematic PPS sampling under substitution of units. *Survey Methodology*, 34, 1, 3-10.

You, Y., Rao, J.N.K. and Dick, P. (2004). Benchmarking hierarchical Bayes small area estimators in the Canadian census undercoverage estimation. *Statistics in Transition*, 6, 631-640.

Survey Methodology, June 2015
Vol. 41, No. 1, pp. 79-97
Statistics Canada, Catalogue No. 12-001-X

Modified regression estimator for repeated business surveys with changing survey frames

John Preston[1]

Abstract

Composite estimation is a technique applicable to repeated surveys with controlled overlap between successive surveys. This paper examines the modified regression estimators that incorporate information from previous time periods into estimates for the current time period. The range of modified regression estimators are extended to the situation of business surveys with survey frames that change over time, due to the addition of "births" and the deletion of "deaths". Since the modified regression estimators can deviate from the generalized regression estimator over time, it is proposed to use a compromise modified regression estimator, a weighted average of the modified regression estimator and the generalised regression estimator. A Monte Carlo simulation study shows that the proposed compromise modified regression estimator leads to significant efficiency gains in both the point-in-time and movement estimates.

Key Words: Changing survey frames; Composite estimation; Modified regression; Repeated surveys; Rotating samples.

1 Introduction

The method of composite estimation has been used extensively in rotating panel household surveys to improve the efficiency of movement estimates, by giving more weight to the "common" rotation groups. Most of the existing composite estimators, such as the AK-composite estimator (Gurney and Daly 1965), the best linear unbiased estimator (BLUE) (Yansaneh and Fuller 1998), and the B1 estimator (Bell 2001), require that all primary sampling units in the population can be assigned randomly to rotation groups. These composite estimators have not been widely adopted for business surveys, as the concept of rotation groups does not translate well to repeated business surveys. Rotating panel designs are not well suited to repeated business surveys due to the highly dynamic nature of the survey frames, with changes caused by the addition of population "births" and the deletion of population "deaths" to the survey frame, as well as changes in classification information on the survey frame over time.

A typical example of this type of repeated business survey is the Quarterly Business Indicators Survey (Australian Bureau of Statistics (ABS) 2012b), where the sampling frame is updated quarterly to take account of new businesses and changes in the characteristics of businesses. Furthermore, approximately one-twelfth of the sampled sector units is rotated out of the survey and is replaced by other units, in order to spread the reporting workload equitably.

The modified regression estimator which was first introduced by Singh (1994) appears to be the most appropriate type of composite estimator suitable to the situation of changing survey frames. The earliest modified regression estimators were the MR1 estimator (Singh and Merkouris 1995; Singh 1996), and the MR2 estimator (Singh, Kennedy, Wu and Brisebois 1997). The former has been found to perform better for point-in-time estimates, while the later has been found to perform better for movement estimates. A compromise between these two modified regression estimators, called the composite modified regression estimator, was suggested by Fuller and Rao (2001). This composite modified regression estimator has

1. John Preston, Australian Bureau of Statistics, 639 Wickham Street, Fortitude Valley QLD 4006, Australia. E-mail: john.preston@abs.gov.au.

been studied by Singh, Kennedy and Wu (2001), Gambino, Kennedy and Singh (2001), Bell (2001) and Beaumont and Bocci (2005).

All of these modified regression estimators perform best when units in the population are unchanged between the previous and current time periods. This will not be too problematic for a typical monthly household survey where the birth, death and net migration rates are relatively low. For example in Australia during 2011-12, the average monthly birth rate was 0.11%, the average monthly death rate was 0.05%, and the average monthly net migration rate was 0.08% (ABS 2012a). However, it will be more problematic for a typical quarterly business survey where the birth and death rates are much larger. For example in Australia during 2011-12, the average quarterly birth or entry rate of businesses was 3.38% and the average quarterly death or exit rate of businesses was 3.28% (ABS 2012c).

If there are significant changes in the population over time, then these modified regression estimators will be unsuitable in their present form, as these estimators can accrue serious biases over time. These modified regression estimators can be extended to the situation of changing survey frames by making adjustments to the composite auxiliary variables, after first adding "births" into the population at the previous time period, and adding "deaths" into the population at the current time period to create a "pseudo-population". These "pseudo-populations" will satisfy the requirement that the units in the population remain unchanged between the previous and current time periods.

Section 2 describes the generalised regression estimator and modified regression estimators, as well as a weighted average of these two estimators which leads to significant efficiency gains in both the point-in-time and movement estimates. An extension to the modified regression estimator for changing survey frames is also outlined in Section 3. The findings of a simulation study are reported in Section 4. Some concluding remarks are provided in Section 5.

2 Modified regression estimation

Consider a finite population $U^{(t)}$ at time t partitioned into H non-overlapping strata $U_1^{(t)},\ldots,U_h^{(t)},\ldots,U_H^{(t)}$, where $U_h^{(t)}$ is comprised of $N_h^{(t)}$ units. A simple random sample without replacement $s_h^{(t)}$ of $n_h^{(t)}$ units is selected with inclusion probabilities $\pi_i^{(t)} = n_h^{(t)} / N_h^{(t)} \left(i \in U_h^{(t)}\right)$ within each stratum h at time t, leading to a total sample $s^{(t)} = \bigcup_{h=1}^{H} s_h^{(t)}$ of size $n^{(t)} = \sum_{h=1}^{H} n_h^{(t)}$. An unbiased estimate of the population total $Y^{(t)} = \sum_{h=1}^{H} \sum_{i \in U_h^{(t)}} y_i^{(t)}$ is given by the Horvitz-Thompson (HT) estimator $\hat{Y}_{\mathrm{HT}}^{(t)} = \sum_{h=1}^{H} \sum_{i \in s_h^{(t)}} w_i^{(t)} y_i^{(t)}$, where $w_i^{(t)} = 1/\pi_i^{(t)}$ is the design weight for unit i at time t and $y_i^{(t)}$ is the value for the variable of interest y for unit i at time t. Assume that there exists a set of auxiliary variables $\mathbf{x}^{(t)}$ at time t for which the population totals $\mathbf{X}^{(t)} = \sum_{i \in U^{(t)}} \mathbf{x}_i^{(t)}$ are known and $\mathbf{x}_i^{(t)}$ are known for every $i \in s^{(t)}$.

The generalised regression (GR) estimator (Särndal, Swensson and Wretman 1992) is a model assisted estimator, designed to improve the accuracy of the estimates by using auxiliary variables that are correlated with the variable of interest. The GR estimator is given by:

$$\hat{Y}_{\mathrm{GR}}^{(t)} = \hat{Y}_{\mathrm{HT}}^{(t)} + \left(\mathbf{X}^{(t)} - \hat{\mathbf{X}}_{\mathrm{HT}}^{(t)}\right)^T \hat{\boldsymbol{\beta}}_{\mathrm{GR}}^{(t)} \tag{2.1}$$

where $\hat{\boldsymbol{\beta}}_{\mathrm{GR}}^{(t)}$ is the vector of linear regression model parameters given by:

$$\hat{\boldsymbol{\beta}}_{\mathrm{GR}}^{(t)} = \left(\sum_{i \in s^{(t)}} \frac{w_i^{(t)} \mathbf{x}_i^{(t)} \mathbf{x}_i^{(t)T}}{c_i^{(t)}} \right)^{-1} \left(\sum_{i \in s^{(t)}} \frac{w_i^{(t)} \mathbf{x}_i^{(t)} y_i^{(t)}}{c_i^{(t)}} \right) \tag{2.2}$$

and $c_i^{(t)}$ are specified factors that relate to the variance structure of the linear regression model associated with the GR estimator $y_i^{(t)} = \mathbf{x}_i^{(t)T} \hat{\boldsymbol{\beta}}_{\mathrm{GR}}^{(t)} + \varepsilon_i^{(t)}$, with $E\left(\varepsilon_i^{(t)}\right) = 0$, $\mathrm{Var}\left(\varepsilon_i^{(t)}\right) = c_i^{(t)} \sigma^2$ and $\mathrm{Cov}\left(\varepsilon_i^{(t)}, \varepsilon_j^{(t)}\right) = 0$ for all $i \neq j$. The GR estimator can also be written as:

$$\hat{Y}_{\mathrm{GR}}^{(t)} = \sum_{i \in s^{(t)}} \tilde{w}_i^{(t)} y_i^{(t)} \tag{2.3}$$

where $\tilde{w}_i^{(t)} = w_i^{(t)} \tilde{g}_i^{(t)}$ and $\tilde{g}_i^{(t)}$ is the $g-$weight for unit i at time t given by:

$$\tilde{g}_i^{(t)} = 1 + \left(\mathbf{X}^{(t)} - \hat{\mathbf{X}}_{\mathrm{HT}}^{(t)} \right)^T \left(\sum_{i \in s^{(t)}} \frac{w_i^{(t)} \mathbf{x}_i^{(t)} \mathbf{x}_i^{(t)T}}{c_i^{(t)}} \right)^{-1} \frac{\mathbf{x}_i^{(t)}}{c_i^{(t)}}. \tag{2.4}$$

At time $t > 1$ define a set of composite auxiliary variables $\mathbf{z}^{(t)}$ for which "pseudo-benchmark", totals $\tilde{\mathbf{Z}}^{(t)}$ (based on the key survey estimates at time $t-1$) are known and $\mathbf{z}_i^{(t)}$ can be derived for every $i \in s^{(t)}$. The modified regression (MR) estimator is the GR estimator where the variables in the regression model are the auxiliary variables $\mathbf{x}^{(t)}$ and the composite auxiliary variables $\mathbf{z}^{(t)}$. The MR estimator given by:

$$\hat{Y}_{\mathrm{MR}}^{(t)} = \hat{Y}_{\mathrm{HT}}^{(t)} + \left(\left(\mathbf{X}^{(t)}, \tilde{\mathbf{Z}}^{(t)} \right) - \left(\hat{\mathbf{X}}_{\mathrm{HT}}^{(t)}, \hat{\mathbf{Z}}_{\mathrm{HT}}^{(t)} \right) \right)^T \hat{\boldsymbol{\beta}}_{\mathrm{MR}}^{(t)} \tag{2.5}$$

where $\hat{\boldsymbol{\beta}}_{\mathrm{MR}}^{(t)}$ is the vector of linear regression model parameters given by:

$$\hat{\boldsymbol{\beta}}_{\mathrm{MR}}^{(t)} = \left(\sum_{i \in s^{(t)}} \frac{w_i^{(t)} \left(\mathbf{x}_i^{(t)}, \mathbf{z}_i^{(t)} \right) \left(\mathbf{x}_i^{(t)}, \mathbf{z}_i^{(t)} \right)^T}{c_i^{(t)}} \right)^{-1} \left(\sum_{i \in s^{(t)}} \frac{w_i^{(t)} \left(\mathbf{x}_i^{(t)}, \mathbf{z}_i^{(t)} \right) y_i^{(t)}}{c_i^{(t)}} \right). \tag{2.6}$$

The MR estimator can also be written as:

$$\hat{Y}_{\mathrm{MR}}^{(t)} = \sum_{i \in s^{(t)}} \breve{w}_i^{(t)} y_i^{(t)} \tag{2.7}$$

where $\breve{w}_i^{(t)} = w_i^{(t)} \breve{g}_i^{(t)}$ and $\breve{g}_i^{(t)}$ is the $g-$weight for unit i at time t given by:

$$\begin{aligned} \breve{g}_i^{(t)} &= 1 + \left(\left(\mathbf{X}^{(t)}, \tilde{\mathbf{Z}}^{(t)} \right) - \left(\hat{\mathbf{X}}_{\mathrm{HT}}^{(t)}, \hat{\mathbf{Z}}_{\mathrm{HT}}^{(t)} \right) \right)^T \\ &\times \left(\sum_{i \in s^{(t)}} \frac{w_i^{(t)} \left(\mathbf{x}_i^{(t)}, \mathbf{z}_i^{(t)} \right) \left(\mathbf{x}_i^{(t)}, \mathbf{z}_i^{(t)} \right)^T}{c_i^{(t)}} \right)^{-1} \frac{\left(\mathbf{x}_i^{(t)}, \mathbf{z}_i^{(t)} \right)}{c_i^{(t)}}. \end{aligned} \tag{2.8}$$

The key to the effectiveness of MR estimator is the definition of the composite auxiliary variables. Ideally, the values for the composite auxiliary variables at time t, would be equal to the values for the key survey variables at time $t-1$. However, due to the rotation of units into and out of sample from one time period to the next, values for the key survey variables at time $t-1$ will be missing by design for those units in the sample at time t which were not in the sample at time $t-1$.

There are several possible techniques available to define the composite auxiliary variables. The earliest modified regression estimators were the MR1 estimator (Singh and Merkouris 1995; Singh 1996), and the MR2 estimator (Singh, Kennedy, Wu and Brisebois 1997) which used values for the composite auxiliary variables given respectively by:

$$\mathbf{z}_{(\mathrm{MR}1)i}^{(t)} = \begin{cases} \mathbf{y}_i^{(t-1)}, & \text{if } i \in s_h^{(t)} \cap s_h^{(t-1)} \\ \bar{\mathbf{Y}}_{(\mathrm{MR})h}^{(t-1)}, & \text{if } i \in s_h^{(t)} \setminus s_h^{(t-1)} \end{cases} \tag{2.9}$$

$$\mathbf{z}_{(\mathrm{MR}2)i}^{(t)} = \begin{cases} \mathbf{y}_i^{(t)} + \left(\sum_{i \in s_h^{(t)}} w_i^{(t)} \Big/ \sum_{i \in s_h^{(t)} \cap s_h^{(t)-1}} w_i^{(t)}\right)\left(\mathbf{y}_i^{(t-1)} - \mathbf{y}_i^{(t)}\right), & \text{if } i \in s_h^{(t)} \cap s_h^{(t-1)} \\ \mathbf{y}_i^{(t)}, & \text{if } i \in s_h^{(t)} \setminus s_h^{(t)-1} \end{cases} \tag{2.10}$$

and $\bar{\mathbf{Y}}_{(\mathrm{MR})h}^{(t-1)}$ are the composite regression estimators of the population mean in stratum h for key survey variables at time $t-1$.

The MR1 values for the composite auxiliary variables use a mean imputation method to impute for the missing values, while the MR2 values use a reverse historical imputation method to impute for the missing values and then modify the non-imputed values so that the HT estimator of the composite auxiliary variables $\hat{\mathbf{Z}}_{\mathrm{HT}}^{(t)} = \sum_{h=1}^{H} \sum_{i \in s_h^{(t)}} w_i^{(t)} \mathbf{z}_{(\mathrm{MR}2)i}^{(t)}$ at time t is unbiased for the corresponding key survey variables $\mathbf{Y}^{(t-1)}$ at time $t-1$.

The MR1 estimator has been found to perform better for point-in-time estimates, while the MR2 estimator has been found to perform better for movement estimates. Fuller and Rao (2001) proposed an alternative estimator that provides a compromise between improving point-in-time estimates and improving movement estimates by using values for the composite auxiliary variables given by:

$$\mathbf{z}_{(\mathrm{MR})i}^{(t)} = (1-\alpha)\,\mathbf{z}_{(\mathrm{MR}1)i}^{(t)} + \alpha \mathbf{z}_{(\mathrm{MR}2)i}^{(t)}. \tag{2.11}$$

The composite auxiliary variable (2.11) requires a decision on the choice of α, which will depend on the correlations over time for the key survey variables, and the relative importance of the point-in-time and movement estimates.

Beaumont and Bocci (2005) proposed a refinement to the composite auxiliary variable, which they proffered did not require an arbitrary choice of α :

$$\mathbf{z}_{(\mathrm{MRR})i}^{(t)} = \begin{cases} \mathbf{y}_i^{(t-1)}, & \text{if } i \in s_h^{(t)} \cap s_h^{(t-1)} \\ \mathbf{y}_i^{(t)} + \left(\sum_{i \in s_h^{(t)} \cap s_h^{(t-1)}} w_i^{(t)} \left(\mathbf{y}_i^{(t-1)} - \mathbf{y}_i^{(t)}\right) \Big/ \sum_{i \in s_h^{(t)} \cap s_h^{(t-1)}} w_i^{(t)}\right), & \text{if } i \in s_h^{(t)} \setminus s_h^{(t-1)}. \end{cases} \tag{2.12}$$

The MRR values for the composite auxiliary variables use a reverse historical imputation method to impute for the missing values and then modify the imputed values so that the HT estimator of the composite auxiliary variables $\hat{\mathbf{Z}}_{\mathrm{HT}}^{(t)} = \sum_{h=1}^{H} \sum_{i \in s_h^{(t)}} w_i^{(t)} \mathbf{z}_{(\mathrm{MRR})i}^{(t)}$ at time t is unbiased for the corresponding key survey variables $\mathbf{Y}^{(t-1)}$ at time $t-1$.

The MR estimators can deviate from the GR estimator over time (Fuller and Rao 2001). In a repeated survey this "drift" problem will be characterized by a substantial deviation which extends over time between the MR estimator and the GR estimators, while in a simulation study it will be characterized by a reduction over time in the relative efficiency of the MR estimator compared to the GR estimators. A potential solution to the "drift" problem would be to use a weighted average of the MR estimator and the GR estimator (Bell 1999) given by:

$$\hat{Y}^{(t)}_{\text{MRC}} = \alpha \hat{Y}^{(t)}_{\text{GR}} + (1 - \alpha)\hat{Y}^{(t)}_{\text{MR}}. \tag{2.13}$$

The compromise modified regression (MRC) estimator should also provide a compromise between the efficiency gains in the point-in-time and movement estimates, as the MR estimators will generally perform better than the GR estimator for movement estimates, but will not always perform better for point-in-time estimates; in particular the MR2 and MRR estimators.

The MRC estimator requires a decision on the choice of α. Using linearization (or Taylor series) methods to approximate the variance of (2.13), a relatively straight forward expression for α can be found which minimises the variance on the movement estimates while maintaining the variance on the point-in-time estimates produced using GR estimator.

The current MR estimators perform best when units in the population are unchanged between the previous and current time periods. If there are significant changes in the population over time, then these modified regression estimators will be unsuitable in their present form, as these estimators can accrue serious biases over time. While a simple factor $\left(\sum_{i\in s_h^{(t-1)}} w_i^{(t-1)} \Big/ \sum_{i\in s_h^{(t)}} w_i^{(t)}\right)$ could be applied to the MR1, MR2 and MRR values to account for the changes in the population size in stratum h between time $t-1$ and time t, these modified regression estimators still can accrue considerable biases over time.

3 Modified regression estimation for changing survey frames

The MR estimators can be extended to the situation of changing survey frames by adding "births" into the population at the previous time period, and adding "deaths" into the population at the current time period to create a "pseudo-population" (Diagram 3.1). These "pseudo-populations" will satisfy the requirement that the units in the population remain unchanged between the previous and current time periods. A full description of the extension to the MR estimator for changing survey frames is outlined below.

Consider a dynamic population which changes over time due to the addition of "births" and the deletion of "deaths". At time t, the union of $U_h^{(t)}$ and $U_h^{(t-1)}$ can be divided into three components. The first component consists of units in the population in stratum h at time $t-1$ but not at time t, referred to as the "death" population $U_{dh}^{(t-1)}$ in stratum h, comprised of $N_{dh}^{(t-1)}$ units. The second component consists of units in the population in stratum h at time $t-1$ and time t, referred to as the "common" population $U_{ch}^{(t-1)} = U_{ch}^{(t)}$ in stratum h, comprised of $N_{ch}^{(t-1)} = N_{ch}^{(t)}$ units. The third component consists of units in the population in stratum h at time t but not at time $t-1$, referred to as the "birth" population $U_{bh}^{(t)}$ in stratum h, comprised of $N_{bh}^{(t)}$ units. Those units in the population which change stratum between time $t-1$ and t are included in the "death" population $U_{dh}^{(t-1)}$ under their stratum at time $t-1$ and are also included in the "birth" population $U_{bh}^{(t)}$ under their stratum at time t.

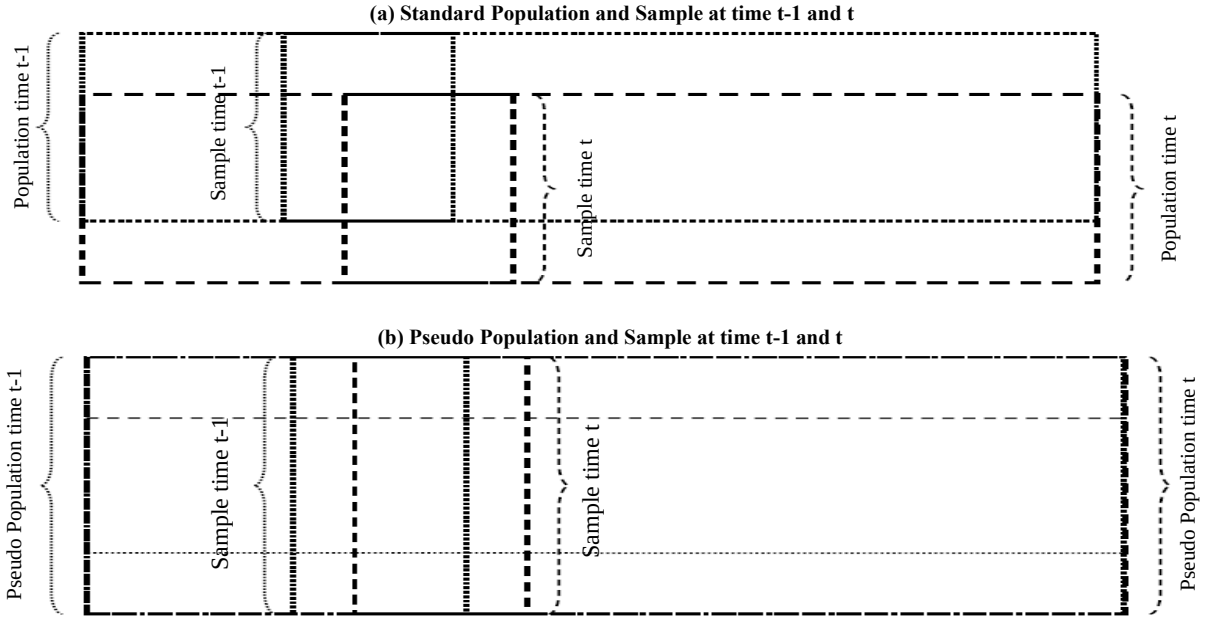


Diagram 3.1 Standard and pseudo populations and samples

At time $t > 1$ define the "pseudo-population" $U_h^{*(t-1)} = U_h^{*(t)}$ in stratum h as the union of $U_h^{(t)}$ and $U_h^{(t-1)}$, comprised of $N_h^{*(t-1)} = N_h^{*(t)} = N_{dh}^{(t-1)} + N_{ch}^{(t-1)} + N_{bh}^{(t)}$ units. It is important to note that the "pseudo-population" $U_h^{*(t-1)}$ at time $t-1$ is different to the "pseudo-population" $U_h^{*(t-1)}$ at time t, as the "pseudo-population" $U_h^{*(t-1)}$ at time $t-1$ is based on the union of $U_h^{(t-2)}$ and $U_h^{(t-1)}$, while the "pseudo-population" $U_h^{*(t-1)}$ at time t is based on the union of $U_h^{(t-1)}$ and $U_h^{(t)}$. Hence the "pseudo-populations" for the current and previous time periods need to be calculated at each time period. Define the "pseudo-values" for the variable of interest y for unit i at time $t-1$ and time t as:

$$y_i^{*(t-1)} = \begin{cases} y_i^{(t-1)}, & \text{if } i \in U_{ch}^{(t-1)} \\ 0, & \text{if } i \in U_{bh}^{(t)} \end{cases}$$

$$y_i^{*(t)} = \begin{cases} y_i^{(t)}, & \text{if } i \in U_{ch}^{(t)} \\ 0, & \text{if } i \in U_{dh}^{(t-1)} \end{cases}$$

and define the "pseudo-values" for the auxiliary variables $\mathbf{x}$ for unit i at time $t-1$ and time t as:

$$\mathbf{x}_i^{*(t-1)} = \begin{cases} \mathbf{x}_i^{(t-1)}, & \text{if } i \in U_{ch}^{(t-1)} \\ 0, & \text{if } i \in U_{bh}^{(t)} \end{cases}$$

$$\mathbf{x}_i^{*(t)} = \begin{cases} \mathbf{x}_i^{(t)}, & \text{if } i \in U_{ch}^{(t)} \\ 0, & \text{if } i \in U_{dh}^{(t-1)}. \end{cases}$$

At time $t > 1$ denote $s_h^{*(t-1)}$ and $s_h^{*(t)}$ as the "pseudo-samples" in stratum h, where $s_h^{*(t-1)}$ consists of all units selected in the original sample $s_h^{(t-1)}$ in stratum h at time $t-1$ plus a random sample of units $s_{bh}^{(t)}$ from the "birth" population $U_{bh}^{(t)}$ in stratum h at time t selected with inclusion probabilities $\pi_i^{(t-1)} = n_h^{(t-1)} / N_h^{(t-1)} \left(i \in U_h^{(t-1)}\right)$, and $s_h^{*(t)}$ consists of all units selected in the original sample $s_h^{(t)}$ in stratum h at time t plus a random sample of units $s_{dh}^{(t-1)}$ from the "death" population $U_{dh}^{(t-1)}$ in stratum h at time $t-1$ selected with inclusion probabilities $\pi_i^{(t)} = n_h^{(t)} / N_h^{(t)} \left(i \in U_h^{(t)}\right)$. Let $n_h^{*(t-1)}$ and $n_h^{*(t)}$ denote the sample sizes in the "pseudo-samples" $s_h^{*(t-1)}$ and $s_h^{*(t)}$ respectively. Once again it is important to note that the "pseudo-sample" $s_h^{*(t-1)}$ at time $t-1$ is different to the "pseudo-sample" $s_h^{*(t-1)}$ at time t, as the "pseudo-sample" $s_h^{*(t-1)}$ at time $t-1$ includes a random sample of units from the "birth" population at time $t-1$, while the "pseudo-sample" $s_h^{*(t-1)}$ at time t includes a random sample of units from the "death" population at time $t-1$. Hence the "pseudo-samples" for the current and previous time periods need to be calculated at each time period.

The choice of an appropriate sample selection technique, for the selection of the additional random samples of units from the "birth" and "death" populations, will depend on the sample selection technique used to select the original samples. Many repeated business surveys select their samples using a permanent random number (PRN) selection technique, to enable some control of the rotation of units into and out of sample from one time period to the next. Consider the simplest case where the original samples $s_h^{(t-1)}$ and $s_h^{(t)}$ in stratum h described by $\left\{i \in U_h^{(t-1)} \;\&\; R_i \in \left[S_h^{(t-1)}, E_h^{(t-1)}\right)\right\}$ and $\left\{i \in U_h^{(t)} \;\&\; R_i \in \left[S_h^{(t)}, E_h^{(t)}\right)\right\}$, where $S_h^{(t)}$ and $E_h^{(t)}$ are the selection interval start and end points in stratum h at time t, and R_i is the permanent random number for unit i. In this case the "pseudo-samples" $s_h^{*(t-1)}$ and $s_h^{*(t)}$ in stratum h are described by $\left\{i \in U_h^{*(t-1)} \;\&\; R_i \in \left[S_h^{(t-1)}, E_h^{(t-1)}\right)\right\}$ and $\left\{i \in U_h^{*(t)} \;\&\; R_i \in \left[S_h^{(t)}, E_h^{(t)}\right)\right\}$. This selection technique will give a similar amount of overlap between the samples from the "death" population at time $t-1$ and t and between the samples from the "birth" population at time $t-1$ and t as between the samples from the "common" population at time $t-1$ and t. Clearly the amount of overlap between the samples from the "death" and "birth" populations will affect the behaviour of the estimates and optimising the amount of overlap could be investigated.

Define the "pseudo-design weights" $w_i^{*(t-1)} = 1/\pi_i^{(t-1)}$ for all units in the "pseudo-sample" $s_h^{*(t-1)}$ and $w_i^{*(t)} = 1/\pi_i^{(t)}$ for all units in the "pseudo-sample" $s_h^{*(t)}$. Since the "pseudo-design weights" for the original sampled units are equal to the original design weights and the "pseudo-values" for the variable of interest are equal to zero for the additional sampled units from the "birth" and "death" populations, then the HT estimator $\hat{Y}_{\mathrm{HT}}^{*(t)} = \sum_{h=1}^{H} \sum_{i \in s_h^{*(t)}} w_i^{*(t)} y_i^{*(t)}$ based on the "pseudo-sample", "pseudo-values" and "pseudo-design weights" is equivalent to the HT estimator $\hat{Y}_{\mathrm{HT}}^{(t)} = \sum_{h=1}^{H} \sum_{i \in s_h^{(t)}} w_i^{(t)} y_i^{(t)}$ based on the original sample, original values and original design weights. Hence the inclusion of these additional sampled units into the "pseudo-sample" from the "birth" and "death" populations will not introduce any extra variability into the point-in-time estimates.

The proposed MR estimator for the special case of changing survey frames can be written as:

$$\hat{Y}_{\mathrm{MR}}^{*(t)} = \sum_{i \in s^{*(t)}} \breve{w}_i^{*(t)} y_i^{*(t)} \tag{3.1}$$

where $\breve{w}_i^{*(t)} = w_i^{*(t)} \breve{g}_i^{*(t)}$ and $\breve{g}_i^{*(t)}$ is the "pseudo$-g-$weight" for unit i at time t given by:

$$\begin{aligned} \breve{g}_i^{*(t)} &= 1 + \left(\left(\mathbf{X}^{(t)}, \tilde{\mathbf{Z}}^{*(t)}\right) - \left(\hat{\mathbf{X}}_{\mathrm{HT}}^{(t)}, \hat{\mathbf{Z}}_{\mathrm{HT}}^{*(t)}\right)\right)^T \\ &\times \left(\sum_{i \in s^{*(t)}} \frac{w_i^{*(t)} \left(\mathbf{x}_i^{*(t)}, \mathbf{z}_i^{*(t)}\right)\left(\mathbf{x}_i^{*(t)}, \mathbf{z}_i^{*(t)}\right)^T}{c_i^{(t)}}\right)^{-1} \frac{\left(\mathbf{x}_i^{*(t)}, \mathbf{z}_i^{*(t)}\right)}{c_i^{(t)}} \end{aligned} \tag{3.2}$$

and the MR1, MR2 and MRR values for the "pseudo-composite auxiliary variables" are given by:

$$\mathbf{z}_{(\mathrm{MR}1)i}^{*(t)} = \begin{cases} R_h^{(t-1,t)} \mathbf{y}_i^{*(t-1)}, & \text{if } i \in s_h^{*(t)} \cap s_h^{*(t-1)} \text{ and } s_h^{*(t)} \setminus s_h^{*(t-1)} \neq \varnothing \\ R_h^{(t-1,t)} \left(\dfrac{\sum_{i \in s_h^{(t)}} w_i^{(t)}}{\sum_{i \in s_h^{*(t)} \cap s_h^{(t-1)}} w_i^{*(t)}}\right) \mathbf{y}_i^{*(t-1)}, & \text{if } i \in s_h^{*(t)} \cap s_h^{*(t-1)} \text{ and } s_h^{*(t)} \setminus s_h^{*(t-1)} = \varnothing \\ R_h^{(t-1,t)} \left(\dfrac{\left(\sum_{i \in s_h^{(t)}} w_i^{(t)} - \sum_{i \in s_h^{*(t)} \cap s_h^{(t-1)}} w_i^{*(t)}\right)}{\sum_{i \in s_h^{*(t)} \setminus s_h^{*(t-1)}} w_i^{*(t)}}\right) \bar{\mathbf{Y}}_{(\mathrm{MR})h}^{(t-1)}, & \text{if } i \in s_h^{*(t)} \setminus s_h^{*(t-1)}. \end{cases} \tag{3.3}$$

$$\mathbf{z}_{(\mathrm{MR}2)i}^{*(t)} = \begin{cases} R_h^{(t-1,t)} \left\{\left(\dfrac{\sum_{i \in s_h^{(t)}} w_i^{(t)}}{\sum_{i \in s_h^{*(t)} \cap s_h^{(t-1)}} w_i^{*(t)}}\right) \mathbf{y}_i^{*(t-1)} + \left(1 - \left(\dfrac{\sum_{i \in s_h^{(t)}} w_i^{(t)}}{\sum_{i \in s_h^{(t)} \cap s_h^{*(t-1)}} w_i^{*(t)}}\right)\right) \mathbf{y}_i^{*(t)}\right\}, & \text{if } i \in s_h^{*(t)} \cap s_h^{*(t-1)} \\ R_h^{(t-1,t)} \mathbf{y}_i^{*(t)}, & \text{if } i \in s_h^{*(t)} \setminus s_h^{*(t-1)}. \end{cases} \tag{3.4}$$

$$\mathbf{z}_{(\mathrm{MRR})i}^{*(t)} = \begin{cases} R_h^{(t-1,t)} \mathbf{y}_i^{*(t-1)}, & \text{if } i \in s_h^{*(t)} \cap s_h^{*(t-1)} \text{ and } s_h^{*(t)} \setminus s_h^{*(t-1)} \neq \varnothing \\ R_h^{(t-1,t)} \left(\sum_{i \in s_h^{(t)}} w_i^{(t)} \Big/ \sum_{i \in s_h^{*(t)} \cap s_h^{(t-1)}} w_i^{*(t)}\right) \mathbf{y}_i^{*(t-1)}, & \text{if } i \in s_h^{*(t)} \cap s_h^{*(t-1)} \text{ and } s_h^{*(t)} \setminus s_h^{*(t-1)} = \varnothing \\ R_h^{(t-1,t)} \left\{\mathbf{y}_i^{*(t)} - \left[\left(\sum_{i \in s_h^{(t)} \setminus s_h^{*(t-1)}} w_i^{*(t)} \Big/ \sum_{i \in s_h^{*(t)} \setminus s_h^{*(t-1)}} w_i^{*(t)}\right)\right.\right. \\ \quad \left.\times \left(\sum_{i \in s_h^{(t)} \cap s_h^{*(t-1)}} w_i^{*(t)} \mathbf{y}_i^{*(t)} \Big/ \sum_{i \in s^{(t)} \cap s^{*(t-1)}} w_i^{*(t)}\right)\right] \\ \quad + \left[\left(\left(\sum_{i \in s_h^{(t)}} w_i^{(t)} - \sum_{i \in s_h^{*(t)} \cap s_h^{(t-1)}} w_i^{*(t)}\right) \Big/ \sum_{i \in s_h^{*(t)} \setminus s_h^{*(t-1)}} w_i^{*(t)}\right)\right. \\ \quad \left.\left.\times \left(\sum_{i \in s_h^{*(t)} \cap s_h^{(t-1)}} w_i^{*(t)} \mathbf{y}_i^{*(t-1)} \Big/ \sum_{i \in s_h^{*(t)} \cap s_h^{(t-1)}} w_i^{*(t)}\right)\right]\right\}, & \text{if } i \in s_h^{*(t)} \setminus s_h^{*(t-1)}. \end{cases} \tag{3.5}$$

where $R_h^{(t-1,t)} = \left(\sum_{i \in s_h^{(t-1)}} w_i^{(t-1)} \Big/ \sum_{i \in s_h^{(t)}} w_i^{(t)}\right)$ is a correction factor applied to the MR1, MR2 and MRR values to account for the relative change in the population size in stratum h between time $t-1$ and time t. The other adjustments to the MR2 and MRR values were made to ensure that the HT estimator for the "pseudo-composite auxiliary variables" $\hat{\mathbf{Z}}_{\mathrm{HT}}^{*(t)} = \sum_{h=1}^{H} \sum_{i \in s_h^{*(t)}} w_i^{*(t)} \mathbf{z}_i^{*(t)}$ at time t is unbiased for the

corresponding key survey variables $\mathbf{Y}^{(t-1)}$ at time $t-1$. A simple proof of the unbiasedness of the HT estimator for the "pseudo-composite auxiliary variables" is shown in the Appendix.

The HT estimator $\hat{Y}_{\text{HT}}^{*(t)} = \sum_{h=1}^{H} \sum_{i \in s_h^{*(t)}} w_i^{*(t)} y_i^{*(t)}$ is equivalent to $\hat{Y}_{\text{HT}}^{(t)} = \sum_{h=1}^{H} \sum_{i \in s_h^{(t)}} w_i^{(t)} y_i^{(t)}$ since the "pseudo-values" for the variable of interest are equal to zero for the additional sampled units from the "birth" and "death" populations. Similarly the GR estimator $\hat{Y}_{\text{GR}}^{*(t)} = \sum_{h=1}^{H} \sum_{i \in s_h^{*(t)}} \tilde{w}_i^{*(t)} y_i^{*(t)}$ is equivalent to $\hat{Y}_{\text{GR}}^{(t)} = \sum_{h=1}^{H} \sum_{i \in s_h^{(t)}} \tilde{w}_i^{(t)} y_i^{(t)}$ since the "pseudo-values" for the variable of interest and the auxiliary variables are equal to zero for the additional sampled units from the "birth" and "death" populations. However, the MR estimator $\hat{Y}_{\text{MR}}^{*(t)} = \sum_{h=1}^{H} \sum_{i \in s_h^{*(t)}} \breve{w}_i^{*(t)} y_i^{*(t)}$ is not equivalent to $\hat{Y}_{\text{MR}}^{(t)} = \sum_{h=1}^{H} \sum_{i \in s_h^{(t)}} \breve{w}_i^{(t)} y_i^{(t)}$ since the "pseudo-values" for the composite auxiliary variables are not equal to zero for the additional sampled units from the "birth" and "death" populations.

The proposed procedure of adding "births" into the population at the previous time period and adding "deaths" into the population at the current time period is performed independently at each time period, so there is no accumulation of "births" and "deaths" in the "pseudo-population" over time.

4 Simulation study

A Monte Carlo simulation study was conducted to examine the performance of the proposed composite regression estimator. Ten artificial populations were created for the simulation study. Firstly, a base population (Population I) was generated to resemble the physical appearance of typical monthly business surveys conducted over a five year time period. Secondly, six additional populations (Populations II to VII) were each generated by modifying one of six key characteristics of the base population to help determine whether this particular characteristic had an impact on the performance of the proposed composite regression estimator. Finally, three supplementary populations (Populations VIII to X) were generated to examine the impact of auxiliary variables on the performance of the proposed composite regression estimator. A brief description of the ten artificial populations is given in Table 4.1.

The population totals at time t for the various artificial populations were produced using the time series model:

$$Y^{(t)} = T^{(t)} + \alpha_2 S^{(t)} + \alpha_3 I^{(t)}$$

where $T^{(t)}, S^{(t)}$ and $I^{(t)}$ are the trend, seasonality and irregular components of the time series given by:

$$T^{(t)} = 1{,}000 + 5(t-1) + 50(1 - \cos(\pi(t-1)/18))$$

$$S^{(t)} = 25[\sin(\pi t/6) - \cos(\pi t/6) + \cos(\pi t/3)]$$

$$I^{(t)} = 25\varepsilon^{(t)}$$

with $\alpha_2 = 1$ for all artificial populations, except Population II (high seasonal series) where $\alpha_2 = 4$, and $\alpha_3 = 1$ for all artificial populations, except Population III (high irregular series) where $\alpha_3 = 4$, and $\varepsilon^{(t)} \sim N(0,1)$. The original $(T^{(t)} + S^{(t)} + I^{(t)})$, seasonally adjusted $(T^{(t)} + I^{(t)})$ and trend $(T^{(t)})$ series for the base artificial population are presented in Figure 4.1.

Table 4.1
Description of the artificial populations

Artificial Populations	Population Descriptions
Population I	Base Series
Population II	High Seasonal Series
Population III	High Irregular Series
Population IV	High Population Rotation Series
Population V	High Sample Rotation Series
Population VI	High Unit Variation Series
Population VII	Low Unit Correlation Series
Population VIII	Base Auxiliary Correlation Series
Population IX	High Auxiliary Correlation Series
Population X	Low Auxiliary Correlation Series

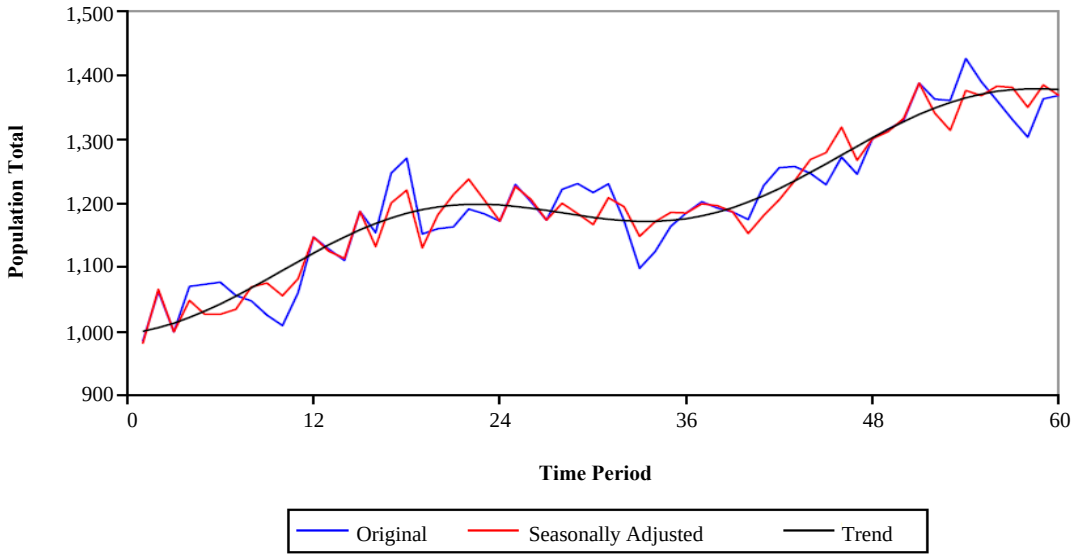


Figure 4.1 Time series for population I

All ten artificial populations were partitioned into five strata; four take-some strata $(h = 1,\ldots,4)$ and one take-all strata $(h = 5)$. The stratum population sizes at time t were chosen as $N_h^{(t)} = N_h\left[1 + 0.5\left(T^{(t)}/T^{(1)} - 1\right)\right]$, where N_h is the stratum population for all artificial populations at time 1, selected to yield a skewed population often associated with typical business.

The expected population rotation rates between time $t-1$ and time t, due to the addition of "births" and the deletion of "deaths", were specified as $\alpha_4(1 - R_h)$, where R_h is the probability of a unit being "deathed" in the population for the base artificial population at any time period. A value of $\alpha_4 = 1$ was used for all artificial populations, except Population IV (high population rotation series) where $\alpha_4 = 2$ was used. The stratum sample sizes at time t were set to $n_h^{(t)} = n_h$ for the take-some strata, and $n_h^{(t)} = N_h^{(t)}$ for the take-all strata, where n_h is the stratum population at time 1.

The planned sample rotation rates between time $t-1$ and time t were specified as $\alpha_5(1 - r_h)$, where r_h is equal to the inverse of the number of consecutive survey cycles a unit is expected to be in the sample given no population rotation, for the base artificial population at any time period (e.g., a planned sample rotation rate of 0.0417 equates to 24 survey cycles). A value of $\alpha_5 = 1$ was used for all artificial

populations, except Population V (high sample rotation series) where $\alpha_5 = 2$ was used. The actual sample rotation rates will depend on these planned sample rotation as well as any unplanned sample rotation caused by the population rotation. The expected population rotation rates and the planned stratum sample rotation rate were selected to yield population and sample rotation rates similar to those often encountered in typical business surveys.

The stratum averages and stratum population variances at time t were specified respectively as $\bar{y}_h^{(t)} = 0.2\left(Y^{(t)}/N_h^{(t)}\right)$ and $S_h^{(t)2} = \alpha_6 S_h^2\left(\bar{y}_h^{(t)}/\bar{y}_h^{(t)}\right)^2$ with $\alpha_6 = 1$ for all artificial populations, except Population VI (high unit variation series) where $\alpha_6 = 4$. The stratum population correlations between time t and time $t-k$ were defined using an exponential decay model, $\rho\left(y_h^{(t)}, y_h^{(t-k)}\right) = \exp\left(-0.02\alpha_7 k\right)$ with $\alpha_7 = 1$ for all artificial populations, except Population VII (low unit correlation series) where $\alpha_7 = 4$. The stratum population correlations between the variable of interest and the auxiliary variable at time t were defined as $\rho\left(x_h^{(t)}, y_h^{(t)}\right) = 1 - \alpha_8\left(1 - \rho_h\right)$ with $\alpha_8 = 1$ for Population VIII (base auxiliary correlation series), $\alpha_8 = 0.5$ for Population IX (high auxiliary correlation series), $\alpha_8 = 1.5$ for Population X (low auxiliary correlation series) and not applicable for all other artificial populations.

The variables of interest $y_{hi}^{(t)}$ and auxiliary variables $x_{hi}^{(t)}$ for unit i in stratum h at time t were generated from multivariate lognormal distributions with means $\bar{y}_h^{(t)}$, variances $S_h^{(t)2}$ and correlation coefficients $\rho\left(y_h^{(t)}, y_h^{(t-k)}\right)$. The stratum level characteristics of N_h, n_h, R_h, r_h and S_h^2 are given by the values presented in Table 4.2.

A total of $S = 10{,}000$ independent simulations were conducted for each of the ten artificial populations. In each of these simulations, stratified random samples $s_h^{(t)}$ of size $n_h^{(t)}$ were selected from the population $U_h^{(t)}$ using a permanent random number (PRN) selection technique at each time period, $t = 1, \ldots, 60$. At each time period, $t > 1$, the "pseudo-populations", $U_h^{*(t-1)}$ and $U_h^{*(t)}$, and "pseudo-samples", $s_h^{*(t-1)}$ and $s_h^{*(t)}$, were identified, and the various MR estimators were evaluated. These included the MR1 estimator $(\alpha = 0)$; the MR2 estimator $(\alpha = 1)$; the MR estimator using $\alpha = 0.25, 0.5$ and 0.75; the MRR estimator and the MRC estimator, with a compromise between the HT estimator and the MRR estimator for Populations I to VII and the GR estimator and the MRR estimator for Populations VIII to X, using $\alpha = 0.25, 0.5$ and 0.75.

Table 4.2
Stratum characteristics

h	N_h	R_h	n_h	r_h	S_h^2	ρ_h
S1	8,000	0.0150	12	0.042	0.4	0.85
S2	1,600	0.0125	18	0.042	3	0.75
S3	320	0.0100	24	0.042	20	0.65
S4	64	0.0075	30	0.000	125	0.55
S5	16	0.0025	16	0.000	625	0.95

The performance of the various MR estimators for the point-in-time and movement estimates were compared using their relative biases and the relative efficiencies with respect to the HT estimator for all artificial populations and also with respect to the GR estimator for Populations VIII to X. The relative biases and relative efficiencies of variable of interest y at time t for the point-in-time and movement estimates were calculated as:

$$\begin{aligned}
\mathrm{RB}\left(\hat{Y}^{(t)}\right) &= \frac{1}{Y^{(t)}}\left[\frac{1}{S}\sum_{s=1}^{S}\left(\hat{Y}_s^{(t)} - Y^{(t)}\right)\right] \\
\mathrm{RB}\left(\hat{Y}^{(t)} - \hat{Y}^{(t-1)}\right) &= \frac{1}{Y^{(t-1)}}\left[\frac{1}{S}\sum_{s=1}^{S}\left(\left(\hat{Y}_s^{(t)} - \hat{Y}_s^{(t-1)}\right) - \left(Y^{(t)} - Y^{(t-1)}\right)\right)\right] \\
\mathrm{RE}\left(\hat{Y}^{(t)}\right) &= \mathrm{MSE}\left(\hat{Y}_*^{(t)}\right)\Big/\mathrm{MSE}\left(\hat{Y}^{(t)}\right) \\
\mathrm{RE}\left(\hat{Y}^{(t)} - \hat{Y}^{(t-1)}\right) &= \mathrm{MSE}\left(\hat{Y}_*^{(t)} - \hat{Y}_*^{(t-1)}\right)\Big/\mathrm{MSE}\left(\hat{Y}^{(t)} - \hat{Y}^{(t-1)}\right)
\end{aligned}$$

where $\hat{Y}_s^{(t)}$ is the estimator for variable of interest y at time t for the s^{th} simulation sample, $\hat{Y}_*^{(t)}$ is the HT or GR estimator for variable of interest y at time t, and $\mathrm{MSE}\left(\hat{Y}^{(t)}\right)$ and $\mathrm{MSE}\left(\hat{Y}^{(t)} - \hat{Y}^{(t-1)}\right)$ are the mean squared errors for variable of interest y at time t for the point-in-time and movement estimates given by:

$$\begin{aligned}
\mathrm{MSE}\left(\hat{Y}^{(t)}\right) &= \frac{1}{S}\sum_{s=1}^{S}\left(\hat{Y}_s^{(t)} - Y^{(t)}\right)^2 \\
\mathrm{MSE}\left(\hat{Y}^{(t)} - \hat{Y}^{(t-1)}\right) &= \frac{1}{S}\sum_{s=1}^{S}\left(\left(\hat{Y}_s^{(t)} - \hat{Y}_s^{(t-1)}\right) - \left(Y^{(t)} - Y^{(t-1)}\right)\right)^2.
\end{aligned}$$

The relative biases of the point-in-time estimates for the MR1, MR2 and MRR estimators, averaged over the twelve months within each of the five years, for Population I (base series) are shown in Table 4.3. The proposed MR estimators (MR1-P, MR2-P, MRR-P) were compared against the current MR estimators (MR1-C, MR2-C, MRR-C), and the adjusted MR estimators (MR1-A, MR2-A, MRR-A), where a correction factor was applied to the MR values to account for the relative change in the population size in stratum h between time $t-1$ and time t.

Table 4.3
Average relative bias (%) of point-in-time estimates for population I

	Year 1	Year 2	Year 3	Year 4	Year 5
HT	0.024	-0.032	-0.015	-0.003	-0.005
MR1-C	-0.909	-2.871	-2.292	-2.836	-4.122
MR2-C	-0.918	-3.432	-3.449	-4.502	-6.820
MRR-C	-0.919	-3.437	-3.458	-4.515	-6.839
MR1-A	0.064	-0.129	0.002	-0.062	-0.068
MR2-A	0.169	0.024	0.039	-0.109	-0.317
MRR-A	0.152	-0.027	-0.014	-0.174	-0.410
MR1-P	0.009	-0.066	-0.040	-0.051	-0.054
MR2-P	0.022	-0.053	-0.028	-0.039	-0.034
MRR-P	0.020	-0.056	-0.030	-0.039	-0.036

The current MR estimators exhibit substantial negative biases which compound over time. While the adjusted MR estimator removes the majority of these biases, the MR2-A and MRR-A estimators still display small negative biases which compound over time. On the other hand, the relative biases of the proposed MR estimator are negligible, with no apparent change in the magnitude of the relative biases over the five years.

Table 4.4 presents the absolute relative biases and relative efficiencies of the estimators for Population I (base series), averaged over the twelve months within each of the five years. The average absolute relative biases of the point-in-time and movement estimates were negligible for all of the estimators, and there was no appreciable change in the magnitude of the relative biases in any of the estimators over the five years. For the point-in-time estimates, the MR1 estimator performed better than the HT estimator, while the MR2 and MRR estimators performed poorer than the HT estimator. The relative efficiency of the MR2 and MRR estimators declined substantially over the five years, which suggests that these estimators are susceptible to the "drift" problem. The presence of the "drift" problem is evident by observing the relationship between the point-in-time estimates at the start of the first year $(t = 1)$ and those at the start of the third year $(t = 25)$ from the simulation samples (Figure 4.2).

It can be seen that there are positive correlations between the point-in-time estimates at the start of the first and third years for the MR1, MR2, MRR and MR $(\alpha = 0.75)$ estimators, signifying that once these estimators vary greatly from the true population totals, then there is a high likelihood that they will continue to drift further from the true population totals over time. While the correlations for the MR1 estimator are lower than those for the MR2 estimator, positive correlations are still evident signifying that the MR1 estimator is not immune from the drift problem. The positive correlations are not apparent for the HT and MRC $(\alpha = 0.25)$ estimators, and hence these estimators are not prone to the "drift" problem. Furthermore, it is clear that the MR2, MRR and MR $(\alpha = 0.75)$ estimators are much more variable than the HT, MR1 and MRC $(\alpha = 0.25)$ estimators at start of the third year.

Table 4.4
Average absolute relative bias (%) and average relative efficiency (%) for population I

	Point-in-Time Estimates					Movement Estimates				
	Year 1	Year 2	Year 3	Year 4	Year 5	Year 1	Year 2	Year 3	Year 4	Year 5
	Average Absolute Relative Bias (%)									
HT	0.031	0.032	0.030	0.025	0.010	0.021	0.011	0.012	0.019	0.014
MR1	0.032	0.066	0.041	0.051	0.054	0.021	0.011	0.010	0.010	0.016
MR2	0.024	0.053	0.030	0.039	0.034	0.014	0.009	0.009	0.009	0.013
MR($\alpha = 0.25$)	0.029	0.067	0.045	0.058	0.063	0.019	0.010	0.009	0.009	0.015
MR($\alpha = 0.50$)	0.027	0.066	0.045	0.060	0.064	0.017	0.010	0.009	0.009	0.014
MR($\alpha = 0.75$)	0.025	0.061	0.040	0.054	0.055	0.016	0.009	0.009	0.009	0.014
MRR	0.023	0.056	0.032	0.040	0.036	0.014	0.009	0.009	0.009	0.013
MRC($\alpha = 0.25$)	0.027	0.041	0.025	0.018	0.011	0.016	0.009	0.010	0.009	0.014
MRC($\alpha = 0.50$)	0.028	0.036	0.028	0.021	0.010	0.018	0.008	0.011	0.010	0.014
MRC($\alpha = 0.75$)	0.029	0.033	0.029	0.024	0.010	0.019	0.008	0.011	0.014	0.014
	Average Relative Efficiency (%)									
HT	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
MR1	122.0	126.0	118.4	112.7	114.6	137.6	132.8	132.7	134.2	133.0
MR2	92.4	74.7	57.7	47.8	45.8	223.0	203.0	206.5	206.4	204.8
MR($\alpha = 0.25$)	121.6	123.4	110.6	100.9	100.9	168.3	158.4	159.7	160.7	159.2
MR($\alpha = 0.50$)	115.3	110.0	92.8	80.9	79.3	199.0	182.8	185.6	186.0	184.3
MR($\alpha = 0.75$)	104.7	91.9	73.5	62.0	59.7	220.4	199.6	203.5	203.4	201.6
MRR	94.1	79.6	63.0	53.4	53.1	223.3	203.3	206.9	206.8	204.8
MRC($\alpha = 0.25$)	110.8	113.7	113.1	113.7	113.1	198.5	182.7	186.5	187.1	184.4
MRC($\alpha = 0.50$)	106.0	105.9	105.6	105.9	105.5	164.2	155.0	157.3	157.6	155.8
MRC($\alpha = 0.75$)	102.7	102.4	102.3	102.4	102.3	130.9	127.4	128.3	128.4	127.6

An appropriate choice of α for the MRC estimators will minimize the likelihood of the "drift" problem. Compared to the MRR estimator, this MRC ($\alpha = 0.25$) estimator will improve the efficiency of the point-in-time estimates, but reduce the efficiency of the movement estimates. For the movement estimates, the MR1 estimator performed slightly better than the HT estimator while the MR2 and MRR estimators performed considerably better than the HT estimator. Overall, the MRC estimator appears to perform slightly better than MR estimator. If the objective is to choose an estimator which is not too susceptible to the "drift" problem and which maximises the efficiency of the movement estimates without any loss in relative efficiency for the point-in-times estimates, then the "best" estimator for this particular population is the MRC estimator with $\alpha \approx 0.10$. This estimator is likely to have minimal drift and leads to moderate efficiency gains of 21.6 percent in the point-in-time estimates and significant efficiency gains of 104.2 percent in the movement estimates.

The average absolute relative biases and average relative efficiencies of the estimators for Populations I to VII are shown in Table 4.5. Large increases in the seasonality (Population II) or irregularity (Population III) of the time series had almost no impact on the performance of the various estimators for the point-in-time estimates. While there were small reductions in the relative efficiency of the movement estimates for MR2 and MRR estimators, there was no impact for the MR1 estimator.

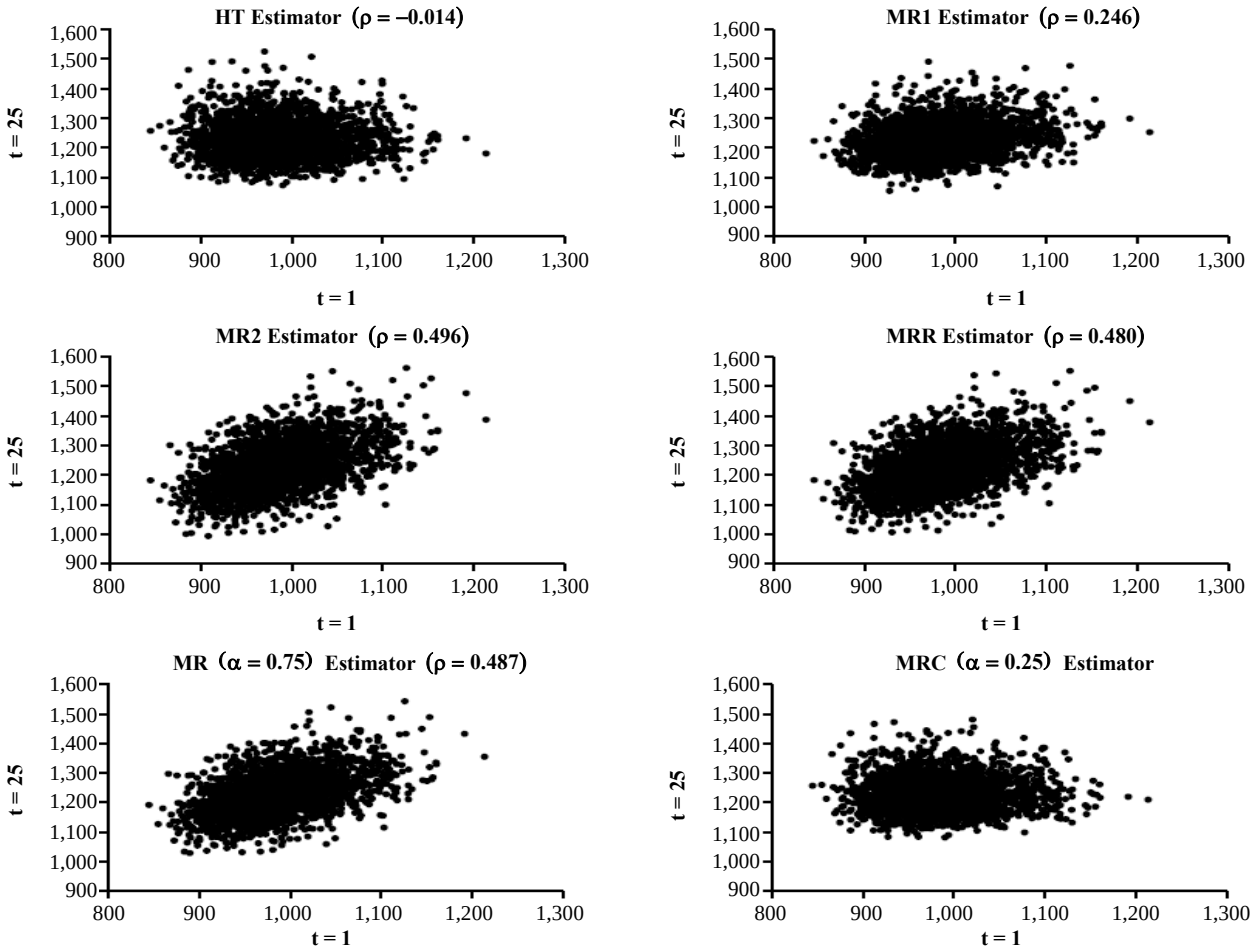


Figure 4.2 Plots of various estimators for population I

Table 4.5
Average absolute relative bias (%) and average relative efficiency (%)

	Point-in-Time Estimates							Movement Estimates						
	Pop I	Pop II	Pop III	Pop IV	Pop V	Pop VI	Pop VII	Pop I	Pop II	Pop III	Pop IV	Pop V	Pop VI	Pop VII
Average Absolute Relative Bias (%)														
HT	0.038	0.027	0.049	0.048	0.048	0.065	0.032	0.017	0.012	0.016	0.018	0.020	0.025	0.020
MR1	0.050	0.098	0.074	0.052	0.089	0.150	0.078	0.014	0.012	0.013	0.015	0.020	0.020	0.018
MR2	0.081	0.028	0.039	0.063	0.047	0.218	0.120	0.012	0.011	0.011	0.014	0.013	0.017	0.017
MR(α = 0.25)	0.052	0.083	0.070	0.046	0.095	0.139	0.090	0.013	0.011	0.012	0.014	0.018	0.018	0.017
MR(α = 0.50)	0.057	0.058	0.059	0.043	0.089	0.136	0.103	0.012	0.010	0.011	0.014	0.016	0.016	0.017
MR(α = 0.75)	0.066	0.038	0.047	0.050	0.069	0.160	0.111	0.012	0.010	0.011	0.014	0.014	0.016	0.017
MRR	0.074	0.032	0.045	0.065	0.055	0.223	0.124	0.012	0.011	0.011	0.014	0.013	0.017	0.017
MRC(α = 0.25)	0.034	0.023	0.046	0.049	0.049	0.059	0.034	0.012	0.010	0.012	0.015	0.015	0.018	0.017
MRC(α = 0.50)	0.037	0.025	0.048	0.049	0.050	0.064	0.033	0.014	0.011	0.014	0.017	0.017	0.023	0.019
MRC(α = 0.75)	0.038	0.026	0.048	0.048	0.049	0.065	0.032	0.015	0.012	0.015	0.018	0.019	0.025	0.019
Average Relative Efficiency (%)														
HT	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
MR1	118.7	119.6	118.9	126.4	143.5	127.2	98.9	134.2	133.4	133.9	132.9	147.2	138.0	115.5
MR2	59.6	60.9	58.1	64.2	49.7	67.8	48.7	208.9	192.6	180.0	202.0	455.7	226.2	137.0
MR(α = 0.25)	110.8	112.0	110.4	119.8	134.2	121.5	89.2	161.6	159.3	158.5	159.0	215.0	169.3	125.7
MR(α = 0.50)	93.6	95.0	92.4	101.4	99.4	103.8	74.6	188.0	182.1	178.2	183.7	315.4	201.0	133.5
MR(α = 0.75)	75.0	76.4	73.5	80.6	69.0	83.8	60.2	206.1	194.9	186.3	200.0	424.9	222.4	137.5
MRR	65.3	66.6	63.7	76.8	52.9	74.0	53.7	209.2	194.6	183.3	202.4	454.8	225.6	137.2
MRC(α = 0.25)	112.9	111.9	112.2	114.5	151.9	112.7	107.5	188.2	183.7	181.4	184.8	347.1	193.7	134.9
MRC(α = 0.50)	105.8	105.4	105.5	107.2	123.3	105.7	104.5	158.3	156.0	154.4	156.6	223.8	160.5	126.2
MRC(α = 0.75)	102.4	102.3	102.3	103.0	109.1	102.4	102.1	128.6	127.9	127.2	128.1	149.7	129.5	114.6

Additional numbers of “births” and “deaths” in the population (Population IV) led to small gains in the relative efficiency of the point-in-time estimates for all of the modified regression estimators, due to reductions in the MSE for the modified regression estimators. While there were small losses in the relative efficiency of the movement estimates for MR2 and MRR estimators, there was no impact for the MR1 estimator. A doubling of the amount of unplanned sample rotation (Population V) produced increases in the relative efficiency of the point-in-time estimates for the MR1 estimator, but decreases in relative efficiency for the MR2 and MRR estimators. There were substantial improvements in relative efficiency of the movement estimates for all of the modified regression estimators as a result of larger increases in the MSE for the HT estimator compared with the modified regression estimators.

Higher unit variation in the reported values (Population VI) led to small gains in the relative efficiency of the point-in-time estimates for all of the modified regression estimators, primarily due to larger increases in the MSE for the HT estimator compared with the modified regression estimators. However, there was no impact in the relative efficiency of the movement estimates as the size of the increases in the MSE for the modified regression estimators were similar to the HT estimator. Low unit correlation in the reported values over time (Population VII) produced large reductions in the relative efficiency of the point-in-time and movement estimates.

Across Populations I to VII, the MR1 estimator performed better than the MR2 and MRR estimators for the point-in-time estimates, while the MR2 and MRR estimators performed better than the MR1 estimator for the movement estimates. The “best” estimator in terms of maximising the relative efficiency

of the movement estimates without any loss in relative efficiency for the point-in-times estimates is the MRC estimator, although the "best" value of α will differ across the different artificial populations.

The average absolute relative biases and average relative efficiencies of the estimators for Populations VIII to X are shown in Table 4.6. With respect to the HT estimator the use of auxiliary variables in the estimators led to large gains in the relative efficiency of the point-in-time estimates and movement estimates for all of the modified regression estimators. The higher the correlation between the variable of interest and the auxiliary variable the greater the gain in relative efficiency of the point-in-time and movement estimates. However, with respect to the GR estimator, the use of auxiliary variables in the estimators led to very small gains in the relative efficiency of the point-in-time estimates, but modest gains in the relative efficiency of the movement estimates for most of the modified regression estimators. The higher the correlation between the variable of interest and the auxiliary variable the lower the gain in relative efficiency of the point-in-time and movement estimates.

Table 4.6
Average absolute relative bias (%) and average relative efficiency (%)

	Point-in-Time Estimates			Movement Estimates		
	Pop VIII	**Pop IX**	**Pop X**	**Pop VIII**	**Pop IX**	**Pop X**
			Average Absolute Relative Bias (%)			
GR	0.021	0.014	0.020	0.010	0.008	0.011
MR1	0.042	0.041	0.044	0.016	0.015	0.016
MR2	0.032	0.026	0.031	0.014	0.013	0.014
MR(α = 0.25)	0.043	0.037	0.044	0.015	0.014	0.015
MR(α = 0.50)	0.041	0.034	0.040	0.015	0.014	0.015
MR(α = 0.75)	0.035	0.029	0.034	0.015	0.013	0.014
MRR	0.036	0.028	0.034	0.014	0.013	0.014
MRC(α = 0.25)	0.023	0.017	0.023	0.013	0.011	0.013
MRC(α = 0.50)	0.022	0.016	0.022	0.012	0.010	0.013
MRC(α = 0.75)	0.021	0.015	0.021	0.011	0.009	0.012
			Average Relative Efficiency (%) to HT Estimator			
GR	256.4	428.9	183.3	169.7	215.3	140.2
MR1	258.9	421.5	191.1	166.8	198.0	150.5
MR2	265.8	436.0	194.4	218.7	247.5	202.2
MR(α = 0.25)	263.8	428.3	194.9	184.4	213.7	168.7
MR(α = 0.50)	267.6	434.7	197.4	202.5	230.5	186.9
MR(α = 0.75)	268.6	438.1	197.3	215.9	244.0	199.8
MRR	266.5	437.5	194.6	216.3	245.8	199.2
MRC(α = 0.25)	266.7	441.2	192.6	225.7	257.7	204.7
MRC(α = 0.50)	265.3	442.0	190.3	217.3	254.4	191.6
MRC(α = 0.75)	261.4	437.0	187.0	197.5	239.7	168.6
			Average Relative Efficiency (%) to GR Estimator			
GR	100.0	100.0	100.0	100.0	100.0	100.0
MR1	101.0	98.3	104.2	98.3	92.0	107.4
MR2	103.7	101.6	106.1	128.9	115.0	144.3
MR(α = 0.25)	102.9	99.9	106.3	108.7	99.3	120.3
MR(α = 0.50)	104.4	101.3	107.7	119.3	107.1	133.3
MR(α = 0.75)	104.8	102.1	107.7	127.2	113.3	142.5
MRR	103.9	102.0	106.1	127.4	114.2	142.1
MRC(α = 0.25)	104.0	102.9	105.1	133.0	119.7	146.0
MRC(α = 0.50)	103.5	103.1	103.8	128.0	118.2	136.7
MRC(α = 0.75)	102.0	101.9	102.0	116.4	111.3	120.3

5 Conclusion

This paper extends a number of the modified regression estimators to business surveys with survey frames that change over time, due to the addition of "births" and the deletion of "deaths". The results of the simulation study indicate that the magnitude of the bias of these various modified regression estimators is negligible. The "best" estimator was the compromise modified regression estimator which led to significant efficiency gains in both the point-in-time and movement estimates, with an appropriate choice of α eliminating the likelihood of the "drift" problem.

Acknowledgements

The views expressed in this paper are those of the author and do not necessarily reflect the views of the Australian Bureau of Statistics (ABS). The author would like to thank the anonymous referees and the associate editor for their valuable comments, and Dr Robert Clark at University of Woollongong for his constructive suggestions on an earlier draft of this manuscript.

Appendix

The expected values of the HT estimator for the "pseudo-composite auxiliary variables" $\hat{\mathbf{Z}}_h^{*(t)} = \sum_{h=1}^{H} \sum_{i \in s_h^{*(t)}} w_i^{*(t)} \mathbf{z}_{(\mathrm{MR}\,2)i}^{*(t)}$ at time t are given by:

$$
\begin{aligned}
E\left[\hat{\mathbf{Z}}_{\mathrm{HT}}^{*(t)}\right] &= E\left[\sum_{h=1}^{H} \sum_{i \in s_h^{*(t)}} w_i^{*(t)} \mathbf{z}_{(\mathrm{MR2})i}^{*(t)}\right] \\
&= \sum_{h=1}^{H} E\left[\sum_{i \in s_h^{*(t)}} w_i^{*(t)} \mathbf{z}_{(\mathrm{MR2})i}^{*(t)}\right] \\
&= \sum_{h=1}^{H} N_h^{(t-1)} E\left[\left(\sum_{i \in s_h^{*(t)} \cap s_h^{*(t-1)}} w_i^{*(t)} \mathbf{y}_i^{*(t-1)} \Big/ \sum_{i \in s_h^{*(t)} \cap s_h^{(t-1)}} w_i^{*(t)}\right)\right] \\
&- \sum_{h=1}^{H} N_h^{(t-1)} E\left[\left(\sum_{i \in s_h^{*(t)} \cap s_h^{*(t-1)}} w_i^{*(t)} \mathbf{y}_i^{*(t)} \Big/ \sum_{i \in s_h^{(t)} \cap s_h^{*(t-1)}} w_i^{*(t)}\right)\right] \\
&+ \sum_{h=1}^{H} N_h^{(t-1)} \left(\sum_{i \in s_h^{(t)} \cap s_h^{*(t-1)}} w_i^{*(t)} \Big/ \sum_{i \in s_h^{(t)}} w_i^{(t)}\right) E\left[\sum_{i \in s_h^{*(t)} \cap s_h^{*(t-1)}} w_i^{*(t)} \mathbf{y}_i^{*(t)} \Big/ \sum_{i \in s_h^{(t)} \cap s_h^{*(t-1)}} w_i^{*(t)}\right] \\
&+ \sum_{h=1}^{H} N_h^{(t-1)} \left(\sum_{i \in s_h^{(t)} \setminus s_h^{*(t-1)}} w_i^{*(t)} \Big/ \sum_{i \in s_h^{(t)}} w_i^{(t)}\right) E\left[\sum_{i \in s_h^{*(t)} \setminus s_h^{*(t-1)}} w_i^{*(t)} \mathbf{y}_i^{*(t)} \Big/ \sum_{i \in s_h^{(t)} \setminus s_h^{*(t-1)}} w_i^{*(t)}\right] \\
&= \sum_{h=1}^{H} N_h^{(t-1)} \bar{\mathbf{Y}}_h^{(t-1)} - \sum_{h=1}^{H} N_h^{(t-1)} \bar{\mathbf{Y}}_h^{(t)} + \sum_{h=1}^{H} N_h^{(t-1)} \bar{\mathbf{Y}}_h^{(t)} \\
&= \sum_{h=1}^{H} N_h^{(t-1)} \bar{\mathbf{Y}}_h^{(t-1)} \\
&= \sum_{h=1}^{H} \mathbf{Y}_h^{(t-1)} \\
&= \mathbf{Y}^{(t-1)}.
\end{aligned}
$$

The expected values of the HT estimator for the "pseudo-composite auxiliary variables" $\hat{\mathbf{Z}}_{\mathrm{HT}}^{*(t)} = \sum_{h=1}^{H} \sum_{i \in s_h^{*(t)}} w_i^{*(t)} \mathbf{z}_{(\mathrm{MRR})i}^{*(t)}$ at time t are given by:

$$\begin{aligned} E\left[\hat{\mathbf{Z}}_{\mathrm{HT}}^{*(t)}\right] &= E\left[\sum_{h=1}^{H} \sum_{i \in s_h^{*(t)}} w_i^{*(t)} \mathbf{z}_{(\mathrm{MRR})i}^{*(t)}\right] \\ &= \sum_{h=1}^{H} E\left[\sum_{i \in s_h^{*(t)}} w_i^{*(t)} \mathbf{z}_{(\mathrm{MRR})i}^{*(t)}\right]. \end{aligned}$$

For the case where there are no units in the "pseudo-sample" in stratum h at time t which were not included in the "pseudo-sample" in stratum h at time $t-1$ $\left(s_h^{*(t)} \setminus s_h^{*(t-1)} = \varnothing\right)$:

$$\begin{aligned} E\left[\sum_{i \in s_h^{*(t)}} w_i^{*(t)} \mathbf{z}_{(\mathrm{MRR})i}^{*(t)}\right] &= N_h^{(t-1)} E\left[\left(\sum_{i \in s_h^{*(t)} \cap s_h^{*(t-1)}} w_i^{*(t)} \mathbf{y}_i^{*(t-1)} \Big/ \sum_{i \in s_h^{*(t)} \cap s_h^{(t-1)}} w_i^{*(t)}\right)\right] \\ &= N_h^{(t-1)} \overline{\mathbf{Y}}_h^{(t-1)} \\ &= \mathbf{Y}_h^{(t-1)} \end{aligned}$$

and for the case where there are units in the "pseudo-sample" in stratum h at time t which were not included in the "pseudo-sample" in stratum h at time $t-1$ $\left(s_h^{*(t)} \setminus s_h^{*(t-1)} \neq \varnothing\right)$:

$$\begin{aligned} E\left[\sum_{i \in s_h^{*(t)}} w_i^{*(t)} \mathbf{z}_{(\mathrm{MRR})i}^{*(t)}\right] &= N_h^{(t-1)} \left(\sum_{i \in s_h^{*(t)} \cap s_h^{(t-1)}} w_i^{*(t)} \Big/ \sum_{i \in s_h^{(t)}} w_i^{(t)}\right) \\ &\times E\left[\sum_{i \in s_h^{*(t)} \cap s_h^{*(t-1)}} w_i^{*(t)} \mathbf{y}_i^{*(t-1)} \Big/ \sum_{i \in s_h^{*(t)} \cap s_h^{(t-1)}} w_i^{*(t)}\right] \\ &+ N_h^{(t-1)} \left(\sum_{i \in s_h^{(t)} \setminus s_h^{*(t-1)}} w_i^{*(t)} \Big/ \sum_{i \in s_h^{(t)}} w_i^{(t)}\right) \\ &\times E\left[\sum_{i \in s_h^{*(t)} \setminus s_h^{*(t-1)}} w_i^{*(t)} \mathbf{y}_i^{*(t)} \Big/ \sum_{i \in s_h^{(t)} \setminus s_h^{*(t-1)}} w_i^{*(t)}\right] \\ &- N_h^{(t-1)} \left(\sum_{i \in s_h^{(t)} \setminus s_h^{*(t-1)}} w_i^{*(t)} \Big/ \sum_{i \in s_h^{(t)}} w_i^{(t)}\right) \\ &\times E\left[\left(\sum_{i \in s_h^{(t)} \cap s_h^{*(t-1)}} w_i^{*(t)} \mathbf{y}_i^{*(t)} \Big/ \sum_{i \in s_h^{(t)} \cap s_h^{*(t-1)}} w_i^{*(t)}\right)\right] \\ &+ N_h^{(t-1)} E\left[\left(\sum_{i \in s_h^{*(t)} \cap s_h^{(t-1)}} w_i^{*(t)} \mathbf{y}_i^{*(t-1)} \Big/ \sum_{i \in s_h^{*(t)} \cap s_h^{(t-1)}} w_i^{*(t)}\right)\right] \\ &- N_h^{(t-1)} \left(\sum_{i \in s_h^{*(t)} \cap s_h^{(t-1)}} w_i^{*(t)} \Big/ \sum_{i \in s_h^{(t)}} w_i^{(t)}\right) \\ &\times E\left[\left(\sum_{i \in s_h^{*(t)} \cap s_h^{(t-1)}} w_i^{*(t)} \mathbf{y}_i^{*(t-1)} \Big/ \sum_{i \in s_h^{*(t)} \cap s_h^{(t-1)}} w_i^{*(t)}\right)\right] \\ &= N_h^{(t-1)} \overline{\mathbf{Y}}_h^{(t-1)} \\ &= \mathbf{Y}_h^{(t-1)} \end{aligned}$$

hence $E\left[\hat{\mathbf{Z}}_{\mathrm{HT}}^{*(t)}\right] = \sum_{h=1}^{H} \mathbf{Y}_h^{(t-1)} = \mathbf{Y}^{(t-1)}$.

References

Australian Bureau of Statistics (ABS) (2012a). Australian demographic statistics, June Quarter 2012, Catalogue Number 3101.0.

Australian Bureau of Statistics (ABS) (2012b). Business indicators, June Quarter 2012, Catalogue Number 5676.0.

Australian Bureau of Statistics (ABS) (2012c). Counts of Australian businesses, including entries and exits, June 2008 to June 2012, Catalogue Number 8165.0.

Beaumont, J.-F., and Bocci, C. (2005). A refinement of the regression composite estimator in the Labour Force Survey for change estimates. *Proceedings of the Survey Methods Section,* SSC Annual Meeting, June 2005.

Bell, P. (1999). Comparison of alternative LFS estimators – Issues for discussion. *Methodology Advisory Committee*, October 1999.

Bell, P. (2001). Comparison of alternative Labour Force Survey estimators. *Survey Methodology*, 27, 1, 53-63.

Fuller, W.A., and Rao, J.N.K. (2001). A regression composite estimator with application to the Canadian Labour Force Survey. *Survey Methodology*, 27, 1, 45-51.

Gambino, J., Kennedy, B. and Singh, M.P. (2001). Regression composite estimation for the Canadian Labour Force Survey: Evaluation and implementation. *Survey Methodology*, 27, 1, 65-74.

Gurney, M., and Daly, J.F. (1965). A multivariate approach to estimation in periodic sample surveys. *Proceedings of the Section of Survey Research Methods*, American Statistical Association, 242-257.

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

Singh, A.C. (1994). Sampling design based estimating functions for finite population totals. Invited paper, *Abstracts of the Annual Meeting of the Statistical Society of Canada*, Banff, Alberta, May 8-11, 48.

Singh, A.C. (1996). Combining information in survey sampling by modified regression. *Proceedings of the Section of Survey Research Methods*, American Statistical Association, 120-129.

Singh, A.C., and Merkouris, P. (1995). Composite estimation by modified regression for repeated surveys. *Proceedings of the Section of Survey Research Methods*, American Statistical Association, 420-425.

Singh, A.C., Kennedy, B. and Wu, S. (2001). Regression composite estimation for the Canadian Labour Force Survey with a rotating panel design. *Survey Methodology*, 27, 1, 33-44.

Singh, A.C., Kennedy, B., Wu, S. and Brisebois, F. (1997). Composite estimation for the Canadian Labour Force Survey. *Proceedings of the Section of Survey Research Methods*, American Statistical Association, 300-305.

Yansaneh, I.S., and Fuller, W.A. (1998). Optimal recursive estimation for repeated surveys. *Survey Methodology*, 24, 1, 31-40.

Survey Methodology, June 2015
Vol. 41, No. 1, pp. 99-126
Statistics Canada, Catalogue No. 12-001-X

Exploring recursion for optimal estimators under cascade rotation

Jan Kowalski and Jacek Wesołowski[1]

Abstract

We are concerned with optimal linear estimation of means on subsequent occasions under sample rotation where evolution of samples in time is designed through a cascade pattern. It has been known since the seminal paper of Patterson (1950) that when the units are not allowed to return to the sample after leaving it for certain period (there are no gaps in the rotation pattern), one step recursion for optimal estimator holds. However, in some important real surveys, e.g., Current Population Survey in the US or Labour Force Survey in many countries in Europe, units return to the sample after being absent in the sample for several occasions (there are gaps in rotation patterns). In such situations difficulty of the question of the form of the recurrence for optimal estimator increases drastically. This issue has not been resolved yet. Instead alternative sub-optimal approaches were developed, as $K-$composite estimation (see e.g., Hansen, Hurwitz, Nisselson and Steinberg (1955)), $AK-$composite estimation (see e.g., Gurney and Daly (1965)) or time series approach (see e.g., Binder and Hidiroglou (1988)).

In the present paper we overcome this long-standing difficulty, that is, we present analytical recursion formulas for the optimal linear estimator of the mean for schemes with gaps in rotation patterns. It is achieved under some technical conditions: ASSUMPTION I and ASSUMPTION II (numerical experiments suggest that these assumptions might be universally satisfied). To attain the goal we develop an algebraic operator approach which allows to reduce the problem of recursion for the optimal linear estimator to two issues: (1) localization of roots (possibly complex) of a polynomial Q_p defined in terms of the rotation pattern (Q_p happens to be conveniently expressed through Chebyshev polynomials of the first kind), (2) rank of a matrix S defined in terms of the rotation pattern and the roots of the polynomial Q_p. In particular, it is shown that the order of the recursion is equal to one plus the size of the largest gap in the rotation pattern. Exact formulas for calculation of the recurrence coefficients are given - of course, to use them one has to check (in many cases, numerically) that ASSUMPTIONs I and II are satisfied. The solution is illustrated through several examples of rotation schemes arising in real surveys.

Key Words: Repeated surveys; Rotation of sample; Recursive BLUE of the current mean; Chebyshev polynomials; Algebra of shift operators; Exponential correlations.

1 Introduction

Repeated surveys with rotation of elements in samples are commonly used by statistical offices and other institutions. Predesigned rotation of (groups of) elements in a form of cascade patterns, that is such schemes when, on each occasion the 'oldest' element (group of elements) leaves the sample and is replaced by a new one, is also very popular but information carried in the survey data is often not exploited in full. This in turn leads to constructing sub-optimal estimators with variance above the achievable minimum. To enhance the use of optimal estimators in rotation schemes, in the seminal paper Patterson (1950) introduced the idea of recurrence for best linear unbiased estimators (BLUEs) of the mean on each occasion. The main assumptions were that the unknown population means are deterministic and the responses are random variables whose variances and correlation structure are fully known. Under exponential correlation and assuming further that any element leaving the sample does not return to the

1. Jan Kowalski, Warsaw University of Technology, Warsaw, Poland; Jacek Wesołowski, Warsaw University of Technology and Central Statistical Office, Warsaw, Poland. E-mail: J.Wesolowski@mini.pw.edu.pl.

survey, Patterson proved that for any occasion t the BLUE $\hat{\mu}_t$ of the current mean μ_t at time t (based on all past observations) can be computed from the following one-step recurrence:

$$\hat{\mu}_t = a_1(t)\hat{\mu}_{t-1} + \underline{r}_0^T(t)\underline{X}_t + \underline{r}_1^T(t)\underline{X}_{t-1} \tag{1.1}$$

where $\underline{X}_i$ is the vector of observations at time $i = t, t-1$. The formulas for the recurrence coefficients, that is the numbers $a_1(t)$ and the vectors $\underline{r}_0(t), \underline{r}_1(t),$ were given there as well. (Here and throughout the paper a vector, say $\underline{r}$, is understood as a column, $\underline{r}^T$ is its transpose. For two vectors $\underline{r} = (r_1, \ldots, r_n)$, $\underline{w} = (w_1, \ldots, w_n) \in \mathbb{R}^n$ the expression $\underline{r}^T\underline{w} = \sum_{i=1}^n r_i w_i$ is just the scalar product of $\underline{r}$ and $\underline{w}$.)

Patterson's assumption that *a unit leaving a sample never returns to the survey* was a core of his approach. If this assumption is violated (that is, there are gaps in rotation patterns) it has been known for years that serious difficulties arise if one seeks an analogue of the recurrence (1.1). Being aware of this (see, e.g., Yansaneh and Fuller 1998) researchers rather tried alternative approaches: Classical K-composite estimator was proposed in Hansen et al. (1955). Its optimality properties were developed in Rao and Graham (1964) and more recently in Ciepiela, Gniado, Wesołowski and Wojtyś (2012). The main difference is that instead of seeking the recurrence for BLUE, these authors restrict the optimality issue to linear unbiased estimators satisfying just the first order recurrence, that is the variance of the estimator based on the most recent estimator and observations from the last two occasions only is minimized. Adjustments, known as AK-composite estimator, introduced in Gurney and Daly (1965), have been developed, e.g., in Cantwell (1988, 1990) and Cantwell and Caldwell (1998) - actually in these papers the authors introduce the notion of balanced multi-level design, and one-level design corresponds to the cascade pattern we consider here. Another approach based on regression composite estimator has been considered in Bell (2001), Fuller and Rao (2001) and Singh, Kennedy and Wu (2001) (with implications for Canadian Labour Force Survey).

The difficulty in recursive estimation in repeated surveys for patterns with gaps was raised in Yansaneh and Fuller (1998), who analyzed variances of composite estimators in several rotation schemes. For a relatively current description of the state of art in the area one can consult Steel and McLaren (2008), in particular Sec. IV on different rotation patterns and Sec. V on composite estimators. Comparisons of effectiveness under different cascade patterns can be found in McLaren and Steel (2000) and Steel and McLaren (2002). A very recent paper on optimal estimation under rotation is by Towhidi and Namazi-Rad (2010). Some of these references deal also with time series approach (which is not considered in this paper) in which the unknown means are treated as random quantities - an overview of such approach can be found in Binder and Hidiroglou (1988). For a more recent development of this setting see e.g., Lind (2005).

As for the original approach of Patterson, the next result concerning the recursive form of the BLUE was presented in Kowalski (2009), where singleton gaps in the rotation pattern were allowed. As in Patterson (1950), this paper was devoted to the "classical" situation in which the coefficients in (1.2) below are allowed to depend on t. Three conclusions from that work have an impact on this paper. Firstly, it was suggested that the formula (1.1) may be generalized to an arbitrary rotation scheme (including gaps in the pattern) by incorporating the optimal estimators and observations from a probably larger (but still as small as possible) number of past occasions and that the order of the recurrence should depend on the size of the largest gap. Secondly, it was observed there that the exponential correlation, as

assumed in Patterson (1950), is crucial for obtaining the recursive representation and that it is plausible to restrict oneself to the class of 'cascade' schemes. Both these assumptions are kept below. Finally, since according to numerical simulations the recurrence coefficients appear to be quickly convergent as $t \to \infty$, a suggestion was made to consider the 'limiting' case of the "classical" setting, in which the recurrence coefficients do not change in time.

We want to stress that in the present paper *any set of gaps in the cascade rotation pattern is allowed*. The aim is to show that the recurrence

$$\hat{\mu}_t = a_1\hat{\mu}_{t-1} + \ldots + a_p\hat{\mu}_{t-p} + \underline{r}_0^T \underline{X}_t + \underline{r}_1^T \underline{X}_{t-1} + \ldots + \underline{r}_p^T \underline{X}_{t-p} \tag{1.2}$$

holds for any cascade rotation scheme and to find the order of recurrence p, the numerical coefficients $a_1, \ldots, a_p$ and the vector coefficients $\underline{r}_0, \ldots, \underline{r}_p$. Let us emphasize that the representation (1.2) is "stationary" in the sense that neither the order of the recurrence p nor the recurrence coefficients (a_i) and $(\underline{r}_i)$ depend on t.

Our main result lies in reducing the recurrence problem to analysis of a certain polynomial Q_p (of degree p, where $p-1$ is the size of the largest gap in the rotation pattern) and to the question of unique solvability of a certain linear system of equations, which depends on roots of Q_p. Luckily the polynomial Q_p happens to be conveniently expressed through Chebyshev polynomials of the first kind. We provide a sufficient condition in terms of localization properties of roots of Q_p for existence of the recursive form of the BLUE of order p, as given in (1.2), and derive explicit formulas (exploiting roots of Q_p) for the recurrence coefficients (a_i) and $(\underline{r}_i)$. The forms of the coefficients depend also on the unique solution of the linear system mentioned above. The result is illustrated by several examples related to the real life surveys.

The convergence of recursion coefficients which we observed numerically in many "classical" schemes (that is, with coefficients in the analogue of (1.2) depending on t) of different complexity indicates that solution to such "stationary" recurrence problem should exist universally (actually only in the Patterson case, $p = 1$, such convergence is formally proved). If so it can be treated as an approximate solution for the "classical" scheme. As the reader will see, this intuition is largely confirmed in this paper. Our main result still is not universal even within models with exponential correlation. Our approach heavily relies on two assumptions (ASSUMPTION I and ASSUMPTION II below) which allow us to claim that the recurrence (1.2) holds true. Nevertheless, we performed many numerical experiments for different rotation patterns and different values of the correlation and they all suggest that both these assumptions may be universally satisfied. Unfortunately, at the present stage we are unable to confirm theoretically these observations.

The plan of the paper is as follows. In Section 2 we introduce in mathematical terms the model we are working with. In Section 3 we introduce our two core assumptions and formulate the main result of the paper. Section 4 contains examples of applications of the main result in several popular rotation schemes. Section 5 presents a discussion. The main body of mathematics is deferred to Section 6. In its first part, 6.1, algebraic properties of shift operators are considered. They are essential for the proof of the recursion formula which is given in the second part, 6.2, of Appendix.

2 Model

Let $\left(X_{i,j}\right)_{i,j\in\mathbb{Z}}$ be a doubly infinite matrix of random variables. Heuristically, $X_{i,j}$ represents the value of variable $\mathcal{X}$ measured for the unit (rotation group) i on the occasion j. We assume that the expectation of $X_{i,j}$ depends only on the occasion and not on the unit, that is

$$\mathbb{E}X_{i,j} = \mu_j, \quad \forall i, j \in \mathbb{Z}.$$

Moreover, we assume exponential in time correlations between $X_{i,j}$'s for the same unit and no correlations between different units (following Patterson (1950) model), that is

$$\mathbb{C}\mathrm{ov}\left(X_{i,j}, X_{k,l}\right) = \rho^{|j-l|}\delta_{i,k} \quad \forall i, j, k, l \in \mathbb{Z},$$

where $|\rho| \in (0,1)$ and $\delta_{i,k} = 1$ if $i = k$, otherwise $\delta_{i,k} = 0$. (In practical situations often ρ is in $[0,1)$. In the case $\rho = 0$ observations from the past cannot improve present linear estimator of the mean, therefore we do not consider such case below.) Consequently,

$$\mathbb{V}\mathrm{ar}\, X_{i,j} = 1, \quad i, j \in \mathbb{Z}.$$

For any $j \in \mathbb{Z}$ we are interested in the BLUE of μ_j based on all available observations from occasions $i \le j$. For a fixed positive integer N denote by

$$\underline{X}_j = \left(X_{j,j}, X_{j+1,j}, \ldots, X_{j+N-1,j}\right)^T$$

the *maximal sample* (of size N) on the occasion $j \in \mathbb{Z}$. Then

$$\mathbb{E}\underline{X}_j = \mu_j \underline{1}, \quad j \in \mathbb{Z},$$

where $\underline{1} = (1,1,\ldots,1)^T \in \mathbb{R}^N$, and

$$\mathbb{C}\mathrm{ov}\left(\underline{X}_j, \underline{X}_{j-k}\right) = \mathbf{C}^k = \left[\mathbb{C}\mathrm{ov}\left(\underline{X}_j, \underline{X}_{j+k}\right)\right]^T, \quad j \in \mathbb{Z}, k \ge 0,$$

where $\mathbf{C}$ is an $N \times N$ matrix of the form

$$\mathbf{C} = \begin{bmatrix} 0 & \rho & & 0 \\ 0 & \ddots & \ddots & \\ & \ddots & \ddots & \rho \\ 0 & & 0 & 0 \end{bmatrix}.$$

Note that $\mathbf{C}^n = 0$ for any $n \ge N$.

The effective sample will be defined by a *cascade pattern*, which is a vector $\underline{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_N)^T \in \{0,1\}^N$ with $\varepsilon_1 = \varepsilon_N = 1$. Let

$$n = \sum_{j=1}^{N} \varepsilon_j \quad \text{and} \quad h = N - n.$$

Let H be the set of zeros in the pattern $\underline{\varepsilon}$, that is $j \in H$ iff $\varepsilon_j = 0$. Obviously, $\#H = h$. A gap of size m is a maximal set of sequential m zeros, that is a set satisfying

$$\{j, j+1, \ldots, j+m-1\} \subset H \quad \text{and} \quad j-1, j+m \notin H.$$

Consequently, H is a union of, say, s gaps of sizes $m_r, r = 1, 2, \ldots, s$, and $\sum_{r=1}^{s} m_r = h$.

The coverage p of the pattern (see Kowalski 2009 for equivalent definition) is the size of the largest gap increased by one:

$$p = 1 + \max_{1 \le r \le s} m_r.$$

On each occasion $j \in \mathbb{Z}$ we may not observe the maximal sample $\underline{X}_j$ but the *effective sample* of size n defined by the cascade pattern $\underline{\varepsilon}$, that is the vector

$$\underline{Y}_j = \left(X_{j+k-1,j}, k \in \{1, \ldots, N\} \setminus H\right)^T,$$

that is values of $X_{i,j}$'s represented by zeros (gaps) in the cascade pattern ε are removed from the sample.

We consider BLUE $\hat{\mu}_t$ of the mean μ_t on the occasion $t \in \mathbb{Z}$ which is based on observations $\underline{Y}_j, j \le t$. That is

$$\hat{\mu}_t = \sum_{i=0}^{\infty} \tilde{\underline{w}}_i^T \underline{Y}_{t-i}$$

with $\tilde{\underline{w}}_i \in \mathbb{R}^n, i \ge 0$, which minimize $\mathbb{V}\text{ar}\hat{\mu}_t$ under the unbiasedness constraints

$$\tilde{\underline{w}}_0^T \underline{1} = 1 \quad \text{and} \quad \tilde{\underline{w}}_i^T \underline{1} = 0, i \ge 1.$$

It is both obvious and crucial for our approach that, equivalently,

$$\hat{\mu}_t = \sum_{i=0}^{\infty} \underline{w}_i^T \underline{X}_{t-i} \tag{2.1}$$

with $\underline{w}_i \in \mathbb{R}^N, i \ge 0$, minimizing $\mathbb{V}\text{ar}\hat{\mu}_t$ under unbiasedness constraints

$$\underline{w}_0^T \underline{1} = 1, \ \underline{w}_i^T \underline{1} = 0, i \ge 1, \tag{2.2}$$

and cascade pattern constraints

$$\underline{w}_i^T \underline{e}_j = 0 \quad \forall i \ge 0, \forall j \in H, \tag{2.3}$$

where $\underline{e}_j = (0, \ldots, 0, 1, 0, \ldots, 0)^T$ (with 1 at j^{th} position) is j^{th} vector of the canonical basis in $\mathbb{R}^N$, $j \in H$. Note that the constraint (2.3) actually says that j^{th} entries $(j \in H)$ of vectors $\underline{w}_i, i \ge 0$, are all zeros.

3 Recurrence

In order to formulate our main result which gives the exact recurrence for the BLUEs under any rotation pattern we need to introduce two objects: a polynomial Q_p and a matrix $\mathbf{S}$. They both look very technical and do not have immediate heuristic interpretations. Nevertheless they appear to be of essential importance for the final recurrence formula.

3.1 Polynomial Q_p

Recall that T_k, the k^{th} Chebyshev polynomial of the first kind, is defined by

$$T_k(x) = \cos(k \arccos x), \quad k = 0,1,\ldots.$$

Define an $m \times m$ symmetric Toeplitz matrix polynomial function $\mathbf{T}_m$ by

$$\mathbf{T}_m = \begin{bmatrix} T_0 & T_1 & T_2 & \cdots & T_{m-2} & T_{m-1} \\ T_1 & T_0 & T_1 & \cdots & T_{m-3} & T_{m-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ T_{m-2} & T_{m-3} & T_{m-4} & \cdots & T_0 & T_1 \\ T_{m-1} & T_{m-2} & T_{m-3} & \cdots & T_1 & T_0 \end{bmatrix} \tag{3.1}$$

and an $m \times m$ tridiagonal invertible matrix

$$\mathbf{R}_m = \begin{bmatrix} 1+\rho^2 & -\rho & 0 & \cdots & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & \cdots & 0 & 0 \\ 0 & -\rho & 1+\rho^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1+\rho^2 & -\rho \\ 0 & 0 & 0 & \cdots & -\rho & 1+\rho^2 \end{bmatrix}. \tag{3.2}$$

Note that $\mathbf{R}_m$ is non-singular.

For a cascade pattern ε with gaps sizes $m_1,\ldots,m_s$ and coverage p define a polynomial Q_p by

$$Q_p(x) = (N-1)\left(1+\rho^2-2\rho x\right) + 1 - \rho^2 - \left(1+\rho^2-2\rho x\right)^2 \sum_{j=1}^{s} \operatorname{tr}\left(\mathbf{T}_{m_j}(x)\,\mathbf{R}_{m_j}^{-1}\right). \tag{3.3}$$

Since $\operatorname{tr}\left(\mathbf{T}_m(x)\,\mathbf{R}_m^{-1}\right)$ is a polynomial of degree $m-1$ in x,

$$\deg Q_p = 2 + \max_{1\le j\le s}\left(m_j - 1\right) = p.$$

3.2 Matrix S

Consider again a cascade pattern $\underline{\varepsilon}$ with coverage p and $\#(H) = h = m_1 + \ldots + m_s$. For complex numbers $d_1, \ldots, d_p$ define a $(ph + h + 1) \times p(h + 1)$ matrix $\mathbf{S}$ through its block structure

$$\mathbf{S} = \mathbf{S}(d_1, \ldots, d_p) = \begin{bmatrix} \widetilde{\mathbf{G}}(d_1) & \widetilde{\mathbf{G}}(d_2) & \cdots & \widetilde{\mathbf{G}}(d_p) \\ \mathbf{G}(d_1) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{G}(d_2) & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{G}(d_p) \end{bmatrix}. \tag{3.4}$$

The blocks $\widetilde{\mathbf{G}}(d_i)$ are $(h + 1) \times (h + 1)$ matrices

$$\widetilde{\mathbf{G}}(d) = \frac{1}{1 - \rho^2} \begin{bmatrix} (N - 1)(1 - d\rho) + 1 - \rho^2 & (1 - d\rho)\, \underline{1}_h^T \\ (1 - d\rho)\, \underline{1}_h & \operatorname{diag}\left(\widetilde{\mathbf{H}}_{m_1}, \ldots, \widetilde{\mathbf{H}}_{m_s}\right) \end{bmatrix} \tag{3.5}$$

with $\widetilde{\mathbf{H}}_m = \widetilde{\mathbf{H}}_m(d)$ being an $m \times m$ upper bi-diagonal matrix

$$\widetilde{\mathbf{H}}_m(d) = \begin{bmatrix} 1 & -d\rho & & \\ & \ddots & \ddots & \\ & & \ddots & -d\rho \\ & & & 1 \end{bmatrix}. \tag{3.6}$$

The blocks $\mathbf{G}(d_i)$ are $h \times (h + 1)$ matrices

$$\mathbf{G}(d) = \frac{1}{1 - \rho^2} \left[(1 - d\rho)(d - \rho)\, \underline{1}_h, d \operatorname{diag}\left(\mathbf{H}_{m_1}, \ldots, \mathbf{H}_{m_s}\right) \right], \tag{3.7}$$

where $\mathbf{H}_m = \mathbf{H}_m(d)$ is an $m \times m$ tri-diagonal matrix

$$\mathbf{H}_m(d) = \begin{bmatrix} 1 + \rho^2 & -d\rho & & \\ -\rho/d & \ddots & \ddots & \\ & \ddots & \ddots & -d\rho \\ & & -\rho/d & 1 + \rho^2 \end{bmatrix}. \tag{3.8}$$

The numbers $d_1, \ldots, d_p$ considered above are related to (potentially complex) roots $x_1, \ldots, x_p$ of the polynomial Q_p through the relation $2x_i = d_i + 1/d_i$, and $|d_i| < 1$, $i = 1, \ldots, p$. Some more details are given in the remark below.

Remark 3.1 *Let $x \in \mathbb{C}$ be such that either $\Im x \neq 0$ or $\Re x \notin [-1, 1]$.*

Then the equation

$$\frac{1}{2}\left(d + \frac{1}{d}\right) = x$$

in d *has exactly two roots, say,* $d_+(x)$ *and* $d_-(x)$ *such that*

$$|d_-(x)| < 1 \quad \textit{and} \quad |d_+(x)| > 1.$$

If additionally $\Im x = 0$ *then* $d_+(x)$ *and* $d_-(x)$ *are real.*

By x^* *denote complex conjugate of* x *with* $\Im x \neq 0$. *Then*

$$d_-(x) = \left(d_-\left(x^*\right)\right)^* \quad \textit{and} \quad d_+(x) = \left(d_+\left(x^*\right)\right)^*.$$

3.3 Main result

Our main result gives the recursion of depth equal to the coverage p of the cascade scheme together with analytic forms of the coefficients which are ready for numerical implementation. Actual examples of such implementations are presented in Section 4. The proof we offer (see Appendix) is based on two basic assumptions concerning the polynomial Q_p and the matrix $\mathbf{S}$.

ASSUMPTION I: The polynomial Q_p has distinct roots $x_1, \ldots, x_p \notin [-1, 1]$.

ASSUMPTION II: The matrix $\mathbf{S} = \mathbf{S}\left(d_1, \ldots, d_p\right)$, where $d_i = d_-(x_i), i = 1, \ldots, p,$ is of full rank.

Theorem 3.1 *If* ASSUMPTIONs I *and* II *are satisfied then for any* $t \in \mathbb{Z}$ *the recursion*

$$\hat{\mu}_t = \sum_{k=1}^{p} a_k \hat{\mu}_{t-k} + \sum_{k=0}^{p} \underline{r}_k^T \underline{X}_{t-k} \tag{3.9}$$

holds with

$$a_k = (-1)^{k+1} \sum_{1 \le j_1 < \ldots < j_k \le p} d_{j_1} \ldots d_{j_k}, \quad k = 1, \ldots, p, \tag{3.10}$$

and

$$\underline{r}_i = \sum_{m=1}^{p} \left[\left(v_i(d_m)\mathbf{I} - v_{i-1}(d_m)\mathbf{C}^T \right) \boldsymbol{\Delta} \mathbf{N}(d_m) \sum_{j \in H'} c_{j,m} \underline{e}_j \right], \quad i = 0, 1, \ldots, p,$$

where $\underline{e}_0 = \underline{1}, H' = \{0\} \cup H, v_0(d) = 1, v_{-1}(d) = 0,$

$$v_i(d) = d^i - \sum_{l=1}^{i} a_l d^{i-l}, \quad i = 1, \ldots, p, \tag{3.11}$$

$\Delta = \left(\mathbf{I} - \mathbf{C}\mathbf{C}^T\right)^{-1}, \mathbf{N}(d) = \mathbf{I} - d\mathbf{C}$ *and with*

$$\underline{c} = \left[\left(c_{j,1},\, j \in H'\right), \left(c_{j,2},\, j \in H'\right), \ldots, \left(c_{j,p},\, j \in H'\right)\right]^T$$

being the unique solution (it exists due to ASSUMPTION II*) of the linear system*

$$\mathbf{S}\underline{c} = (1, 0, \ldots, 0)^T \in \mathbb{R}^{ph+h+1}.$$

Moreover,

$$\mathbb{V}\mathrm{ar}\,(\hat{\mu}_t) = \sum_{m=1}^{p} c_{0,m}. \tag{3.12}$$

In the next section we show how the above theoretical result can be applied in several basic schemes, in particular, in those which are used in real life surveys, while the proof of Theorem 3.1 is given in the second part, 6.2, of Appendix. It is based on a purely algebraic operator approach which is introduced earlier in the first part, 6.1, of Appendix.

We would like to stress that intensive numerical experiments suggest that ASSUMPTIONS I and II may be universally satisfied, however at this moment we do not have mathematical proof of this fact (except the case $p = 1, 2$ and $p = 3$ for a special rotation pattern). Thus applications of the above recursion formula (for $p > 2$) in surveys have to be preceded by a numerical check (which is rather straightforward) that ASSUMPTIONS I and II are satisfied. Examples are given in Section 4.

4 Examples

4.1 Patterson's scheme, $p = 1$

The cascade Patterson scheme is used e.g., for conducting the Labour Force Survey in Australia ($N = n = 8$, see Australian Bureau of Statistics (2002)) and Canada ($N = n = 6$, see Singh, Drew, Gambino and Mayda (1990)). There are no zeros in the pattern, hence $h = 0$ and the polynomial $Q_p = Q_1$, see (3.3), does not contain the summand with the trace, that is

$$Q_1(x) = (N-1)\left(1 + \rho^2 - 2\rho x\right) + 1 - \rho^2.$$

Its only root $x_1 = -\dfrac{1+\rho^2}{2\rho} - \dfrac{1-\rho^2}{2(N-1)\rho}$ is real and satisfies $|x_1| > \dfrac{1+\rho^2}{2|\rho|} > 1$, that is ASSUMPTION I is satisfied. It yields also real $d_1 = d_-(x_1)$ of the form

$$d_1 = \frac{N + (N-2)\rho^2 - \sqrt{\left[N + (N-2)\rho^2\right]^2 - 4(N-1)^2\rho^2}}{2(N-1)\rho}.$$

Moreover, $\mathbf{S}$ as defined in (3.4) is a 1×1 matrix of the form $\mathbf{S} = \left[(N-1)\dfrac{1-d_1\rho}{1-\rho^2} + 1\right] \neq \mathbf{0}$, that is ASSUMPTION II trivially holds. Thus from Theorem 3.1, for all $t \in \mathbb{Z}$ we have

$$\hat{\mu}_t = a_1 \hat{\mu}_{t-1} + \underline{r}_0^T \underline{X}_t + \underline{r}_1^T \underline{X}_{t-1},$$

where

$$\begin{cases} a_1 = d_1 \\ \underline{r}_0 = c_{0,1} \mathbf{N}(d_1) \underline{1} \\ \underline{r}_1 = -c_{0,1} \mathbf{C}^T \mathbf{N}(d_1) \underline{1} \end{cases},$$

where

$$c_{0,1} = \frac{1}{(N-1)\dfrac{1-d_1\rho}{1-\rho^2} + 1}.$$

Taking for example $N = 6$ and $\rho = 0.9$, we obtain for all t :

$$\hat{\mu}_t = 0.7942 \hat{\mu}_{t-1} + \begin{bmatrix} 0.1765 \\ 0.1765 \\ 0.1765 \\ 0.1765 \\ 0.1765 \\ 0.1176 \end{bmatrix}^T \underline{X}_t + \begin{bmatrix} 0.0000 \\ -0.1588 \\ -0.1588 \\ -0.1588 \\ -0.1588 \\ -0.1588 \end{bmatrix}^T \underline{X}_{t-1}.$$

Remark 4.1 *Patterson (1950) considered the same scheme in the "classical" model. The recurrence coefficient $a_1(t)$ was formally proved to converge with $t \to \infty$ and the limit was shown to be a_1 as given above. The vectors $\underline{r}_0(t)$ and $\underline{r}_1(t)$, being continuous functions of $a_1(t)$, converge to $\underline{r}_0$ and $\underline{r}_1$, respectively. That is, the "stationary" solution is indeed consistent with asymptotics of the "classical" one.*

4.2 Schemes with gaps of size 1, $p = 2$

The polynomial $Q_p = Q_2$, see (3.3), has the following form:

$$Q_2(x) = -\frac{4h\rho^2}{1+\rho^2} x^2 - 2(N - 2h - 1)\rho x + (N - h - 1)\left(1 + \rho^2\right) + 1 - \rho^2.$$

As $1 - \rho^2 > 0$, it is immediate that its discriminant

$$\Delta = 4(N - 2h - 1)^2 \rho^2 + 4\frac{4h\rho^2}{1+\rho^2}\left[(N - h - 1)\left(1 + \rho^2\right) + 1 - \rho^2\right] > 4\rho^2 (N-1)^2 > 0. \quad (4.1)$$

Thus Q_2 has two single real roots

$$x_{\pm} = \left(1 + \rho^2\right)\frac{-2\left(N - 2h - 1\right)\rho \pm \sqrt{\Delta}}{8h\rho^2}.$$

Note that since the size of all gaps is one, then necessarily $N - h - 1 \geq h \geq 1$. Using this fact and inequality (4.1) we obtain therefore

$$|x_{\pm}| > \left(1 + \rho^2\right)\frac{N - h - 1}{2|\rho|} \geq \frac{1 + \rho^2}{2|\rho|} > 1, \quad \text{since} \quad |\rho| \in (0,1).$$

Thus the ASSUMPTION I of Theorem 3.1 is satisfied.

By Remark 3.1 it follows that $d_1 = d_-(x_-) = x_- + \sqrt{x_-^2 - 1} < 0$ and $d_2 = d_-(x_+) = x_+ - \sqrt{x_+^2 - 1} > 0$ are real numbers.

Since in this case $s = h$ and $m_1 = \ldots = m_h = 1$ we have $\widetilde{\mathbf{H}}_1(d_i) = 1$ and $\mathbf{H}_1(d_i) = 1 + \rho^2, i = 1,2$. Therefore the equation $\mathbf{S}\underline{c} = \underline{e}$ implies

$$(1 - d_i\rho)(d_i - \rho)c_{0,i} + \left(1 + \rho^2\right)c_{k,i} = 0, \quad k = 1,\ldots,h, \quad i = 1,2.$$

Thus $c_{1,1} = c_{2,1} = \ldots = c_{h,1}$ and $c_{1,2} = c_{2,2} = \ldots = c_{h,2}$. Consequently, the system $\mathbf{S}\underline{c} = \underline{e}$ reduces to the system with four unknowns $c_{0,1}, c_{1,1}, c_{0,2}$ and $c_{1,2}$:

$$\tilde{\mathbf{S}}\left(c_{0,1}, c_{1,1}, c_{0,2}, c_{1,2}\right)^T = (1,0,0,0)^T$$

with

$$\tilde{\mathbf{S}} = \frac{1}{1 - \rho^2}\begin{bmatrix} (N-1)(1 - d_1\rho) + 1 - \rho^2 & h(1 - d_1\rho) & (N-1)(1 - d_2\rho) + 1 - \rho^2 & h(1 - d_2\rho) \\ 1 - d_1\rho & 1 & 1 - d_2\rho & 1 \\ (1 - d_1\rho)(d_1 - \rho) & d_1\left(1 + \rho^2\right) & 0 & 0 \\ 0 & 0 & (1 - d_2\rho)(d_2 - \rho) & d_2\left(1 + \rho^2\right) \end{bmatrix}.$$

To show that $\tilde{\mathbf{S}}$ is non-singular we first show that

$$\rho(d_1 + d_2) \geq 0. \qquad (4.2) \qquad (4.2)$$

To this end we first note that

$$\rho(x_- + x_+) = -\left(1 + \rho^2\right)\frac{N - 2h - 1}{2h} \leq 0. \qquad (4.3)$$

Moreover,

$$\begin{aligned} \rho(d_1 + d_2) &= \rho\left(x_- + x_+\sqrt{x_-^2 - 1} - \sqrt{x_+^2 - 1}\right) = \rho(x_- + x_+)\left(1 + \frac{x_- - x_+}{\sqrt{x_-^2 - 1} + \sqrt{x_+^2 - 1}}\right) \\ &= \frac{\rho(x_- + x_+)}{\sqrt{x_-^2 - 1} + \sqrt{x_+^2 - 1}}\left(\sqrt{x_-^2 - 1} + x_- + \sqrt{x_+^2 - 1} - x_+\right). \end{aligned}$$

Due to (4.3) the last expression is non-negative since the second factor is strictly negative. Now we are ready to consider the determinant

$$\det \tilde{\mathbf{S}} = \frac{(d_2 - d_1)\rho}{(1-\rho^2)^4} s(d_1, d_2),$$

where

$$\begin{aligned} s(d_1, d_2) &= (1+\rho^2)\left[(N-1)(1-d_1\rho)(1-d_2\rho) + (1-\rho^2)(1+d_1d_2\rho^2)\right] \\ &+ h(1-d_1\rho)(1-d_2\rho)(-1+(d_1+d_2)\rho + d_1d_2\rho^2 - 2\rho^2). \end{aligned}$$

We note that $|d_i| < 1, i = 1, 2,$ and thus $|d_1 d_2| < 1$. Consequently, we have $1 + \rho^2 > (1 - d_1\rho)(1 - d_2\rho) > 0, 1 + d_1 d_2 \rho^2 > 0$. These inequalities together with (4.2) yield

$$\begin{aligned} s(d_1, d_2) &> (1-d_1\rho)(1-d_2\rho)\{(N-1)(1+\rho^2) - h[1 + d_1d_2\rho^2 + 2\rho^2]\} \\ &> (1-d_1\rho)(1-d_2\rho)[(N-h-1)(1+\rho^2) - 2h\rho^2] \\ &> (1-d_1\rho)(1-d_2\rho)(N-2h-1)(1+\rho^2) \\ &\geq 0. \end{aligned}$$

Consequently, $\det \tilde{\mathbf{S}} \neq 0$.

Since $\text{rank } \mathbf{S} = \text{rank } \tilde{\mathbf{S}} + 2(h-1)$ we obtain $\text{rank } \mathbf{S} = 2(h+1)$ and thus the ASSUMPTION II of Theorem 3.1 is satisfied. Moreover, $\tilde{\mathbf{S}}^{-1}$ exists. Therefore

$$(c_{0,1}, c_{1,1}, c_{0,2}, c_{1,2}) = (1, 0, 0, 0)\left[\tilde{\mathbf{S}}^{-1}\right]^T.$$

Finally, we conclude that the recurrence has the following form:

$$\hat{\mu}_t = a_1\hat{\mu}_{t-1} + a_2\hat{\mu}_{t-2} + \underline{r}_0^T \underline{X}_t + \underline{r}_1^T \underline{X}_{t-1} + \underline{r}_2^T \underline{X}_{t-2},$$

where

$$\begin{cases} a_1 = d_1 + d_2 \\ a_2 = -d_1 d_2 \\ \underline{r}_0 = \mathbf{N}(d_1)\left[(c_{0,1} + c_{1,1})\underline{1} - c_{1,1}\underline{\varepsilon}\right] + \mathbf{N}(d_2)\left[(c_{0,2} + c_{1,2})\underline{1} - c_{1,2}\underline{\varepsilon}\right] \\ \underline{r}_1 = -(d_2\mathbf{I} + \mathbf{C}^T)\mathbf{N}(d_1)\left[(c_{0,1} + c_{1,1})\underline{1} - c_{1,1}\underline{\varepsilon}\right] - (d_1\mathbf{I} + \mathbf{C}^T)\mathbf{N}(d_2)\left[(c_{0,2} + c_{1,2})\underline{1} - c_{1,2}\underline{\varepsilon}\right] \\ \underline{r}_2 = d_2\mathbf{C}^T\mathbf{N}(d_1)\left[(c_{0,1} + c_{1,1})\underline{1} - c_{1,1}\underline{\varepsilon}\right] + d_1\mathbf{C}^T\mathbf{N}(d_2)\left[(c_{0,2} + c_{1,2})\underline{1} - c_{1,2}\underline{\varepsilon}\right] \end{cases}.$$

For example, let $N = 7, h = 2, H = \{3, 6\}$ and let $\rho = 0.5$. Then

$$Q_2(x) = -1.6x^2 - 2x + 5.75$$

and

$$\begin{cases} x_1 = -2.6211 \\ x_2 = 1.3711 \end{cases} \Rightarrow \begin{cases} d_+(x_1) = -5.0439 \\ d_1 = d_-(x_1) = -0.1983 \\ d_+(x_2) = 2.3091 \\ d_2 = d_-(x_2) = 0.4331 \end{cases} \Rightarrow \begin{cases} a_1 = 0.2348 \\ a_2 = 0.0859 \end{cases}.$$

Finally, (3.9) assumes the form

$$\hat{\mu}_t = 0.2348\hat{\mu}_{t-1} + 0.0859\hat{\mu}_{t-2} + \begin{bmatrix} 0.2171 \\ 0.1904 \\ 0.0000 \\ 0.2171 \\ 0.1904 \\ 0.0000 \\ 0.1850 \end{bmatrix}^T \underline{X}_t + \begin{bmatrix} -0.0093 \\ -0.1086 \\ 0.0000 \\ -0.0093 \\ -0.1086 \\ 0.0000 \\ 0.0010 \end{bmatrix}^T \underline{X}_{t-1} + \begin{bmatrix} 0.0000 \\ 0.0047 \\ 0.0000 \\ -0.0476 \\ 0.0047 \\ 0.0000 \\ -0.0476 \end{bmatrix}^T \underline{X}_{t-2}.$$

4.3 Szarkowski's scheme, $p = 3$

If there are h_2 gaps of size 2 and h_1 gaps of size 1 in the cascade pattern the polynomial $Q_p = Q_3$, see (3.3), assumes the form

$$Q_3(x) = (N-1)\left(1+\rho^2-2\rho x\right) + 1 - \rho^2 - \left(1+\rho^2-2\rho x\right)^2 \left(h_2 \frac{2\rho x + 2\left(1+\rho^2\right)}{1+\rho^2+\rho^4} + h_1 \frac{1}{1+\rho^2} \right).$$

The Szarkowski's scheme is defined by the cascade pattern $\underline{\varepsilon} = (1,1,0,0,1,1)^T$ (often denoted also as $2-2-2$), used e.g., by the Central Statistical Office of Poland for conducting the Labour Force Survey (known under the label BAEL), see Szarkowski and Witkowski (1994) or Popiński (2006). Actually, such scheme is used also in LFS in other countries in Europe as well. Here $N = 6$ and $H = \{3,4\}$. Thus $h_2 = 1, h_1 = 0$, and

$$Q_3(x) = 5\left(1+\rho^2-2\rho x\right) + 1 - \rho^2 - 2\left(1+\rho^2-2\rho x\right)^2 \frac{\rho x + 1 + \rho^2}{1+\rho^2+\rho^4}. \tag{4.4}$$

Wesołowski (2010) proved that in this case Q_3 is either strictly increasing or decreasing in the whole domain and has two complex conjugate roots x_1, x_2, and one real root $x_3 \notin [-1,1]$, meaning that the ASSUMPTION I of Theorem 3.1 holds. It was also shown in that paper that the matrix $\mathbf{S}$, in this case of dimensions 9×9, is invertible (meaning that the ASSUMPTION II of Theorem 3.1 holds). Thus, just as for $p = 1, 2$, the recurrence (3.9) for Szarkowski's scheme always holds.

In general, even in the case $p = 3$, verification of ASSUMPTIONs I and II of Theorem 3.1 has to be done numerically, i.e., after assigning the value to the correlation coefficient ρ. However, it is worth

noting that all performed simulations confirm existence of the solution. Asymptotic approximation of the "classical" model parameters was also observed in numerical experiments we performed.

The coefficients a_1, a_2, a_3 depend on $d_1 = d_-(x_1), d_2 = d_-(x_2) = d_1^*$ and $d_3 = d_-(x_3)$ in the following way (see (3.10)):

$$\begin{cases} a_1 = d_1 + d_2 + d_3 \\ a_2 = -(d_1 d_2 + d_2 d_3 + d_1 d_3). \\ a_3 = d_1 d_2 d_3 \end{cases}$$

For the Szarkowski scheme, taking for instance $\rho = 0.7$ in (4.4), we obtain

$$\begin{cases} x_1 = -0.5668 - 1.4069i \\ x_2 = -0.5668 + 1.4069i \\ x_3 = 1.1336 \end{cases} \Rightarrow \begin{cases} d_+(x_1) = -1.0368 - 3.1035i \\ d_1 = d_-(x_1) = -0.0968 + 0.2899i \\ d_+(x_2) = -1.0368 + 3.1035i \\ d_2 = d_-(x_2) = -0.0968 - 0.2899i \\ d_+(x_3) = 1.6675 \\ d_3 = d_-(x_3) = 0.5997 \end{cases} \Rightarrow \begin{cases} a_1 = 0.4060 \\ a_2 = 0.0227. \\ a_3 = 0.0560 \end{cases}$$

Due to Theorem 3.1 we get the following form of (3.9):

$$\begin{aligned} \hat{\mu}_t &= 0.4060\hat{\mu}_{t-1} + 0.0227\hat{\mu}_{t-2} + 0.0560\hat{\mu}_{t-3} \\ &+ \begin{bmatrix} 0.2862 \\ 0.2217 \\ 0.0000 \\ 0.0000 \\ 0.2862 \\ 0.2059 \end{bmatrix}^T \underline{X}_t + \begin{bmatrix} -0.0036 \\ -0.2004 \\ 0.0000 \\ 0.0000 \\ -0.0036 \\ -0.1984 \end{bmatrix}^T \underline{X}_{t-1} + \begin{bmatrix} -0.0143 \\ 0.0026 \\ 0.0000 \\ 0.0000 \\ -0.0143 \\ 0.0033 \end{bmatrix}^T \underline{X}_{t-2} + \begin{bmatrix} 0.0000 \\ 0.0100 \\ 0.0000 \\ 0.0000 \\ -0.0760 \\ 0.0100 \end{bmatrix}^T \underline{X}_{t-3}. \end{aligned}$$

4.4 CPS scheme, $p = 9$

Let us consider the well-known and widely studied 4-8-4 scheme, that is the cascade pattern is

$$\underline{\varepsilon} = (1,1,1,1,0,0,0,0,0,0,0,0,1,1,1,1)^T$$

which is used in the US in the Current Population Survey, see U.S. Bureau of Census (2002). In this case $N = 16, h = 8,$ and $H = \{5,\ldots,12\}$. We do not have any analytical proof that ASSUMPTIONs I and II are satisfied in this scheme for any ρ.

The polynomial $Q_p = Q_9$, see (3.3), is of degree 9 and has the form

$$Q_9(x) = 15\left(1+\rho^2-2\rho x\right) + 1 - \rho^2 - \left(1+\rho^2-2\rho x\right)^2 \operatorname{tr}\left(\mathbf{T}_8(x)\,\mathbf{R}_8^{-1}\right).$$

Consequently, its analysis, as well as analysis of matrix $\mathbf{S}$ (which is of dimension 81×81 in this scheme), can be done numerically, after assigning some value for ρ. To make use of the result of Theorem 3.1 we need to check numerically that ASSUMPTIONs I and II are satisfied for a given concrete value for ρ. We checked that they hold true for several values for ρ picked up at random from the interval $(-1,1)$.

Taking for instance $\rho = 0.9$, we obtain that Q_9 has eight complex roots and one real root of the form

$$\begin{cases} x_1 = -0.7667 - 0.0208i \\ x_2 = -0.7667 + 0.0208i \\ x_3 = -0.1746 - 0.0320i \\ x_4 = -0.1746 + 0.0320i \\ x_5 = 0.4989 - 0.0284i \\ x_6 = 0.4989 + 0.0284i \\ x_7 = 0.9391 - 0.0121i \\ x_8 = 0.9391 + 0.0121i \\ x_9 = -1.0006 \end{cases} \Rightarrow \begin{cases} d_1 = d_-(x_1) = -0.7419 - 0.6220i \\ d_2 = d_-(x_2) = -0.7419 + 0.6220i \\ d_3 = d_-(x_3) = -0.1689 - 0.9532i \\ d_4 = d_-(x_4) = -0.1689 + 0.9532i \\ d_5 = d_-(x_5) = 0.4825 - 0.8389i \\ d_6 = d_-(x_6) = 0.4825 + 0.8389i \\ d_7 = d_-(x_7) = 0.9064 - 0.3335i \\ d_8 = d_-(x_8) = 0.9064 + 0.3335i \\ d_9 = d_-(x_9) = -0.9682 \end{cases} \Rightarrow \begin{cases} a_1 = 0.7429 \\ a_2 = 0.0019 \\ a_3 = 0.0023 \\ a_4 = 0.0029 \\ a_5 = 0.0037 \\ a_6 = 0.0049 \\ a_7 = 0.0066 \\ a_8 = 0.0088 \\ a_9 = 0.0119 \end{cases}.$$

The coefficient a_1 is dominant in terms of absolute value. The second largest, a_9 is smaller by one order of magnitude and the other coefficients by at least two. Results for other values of the parameter ρ behave similarly.

5 Discussion

The main result of the paper is an explicit recurrence formula for the best linear unbiased estimator (BLUE) of the mean on any occasion in repeated surveys with any cascade rotation pattern. The principal novelty lies in allowing for gaps in the pattern. The results which have been known earlier either dealt with patterns with no gaps or with estimators which were not BLUEs. The approach, we developed, is heavily based on algebra of matrices and linear operators of infinite dimension as well as on properties of Chebyshev polynomials. Unfortunately, the explicit recursive formula we obtained in Theorem 3.1 needs two, seemingly technical, assumptions: ASSUMPTION I on localization of roots of a polynomial Q_p and ASSUMPTION II on rank of matrix $\mathbf{S}$. It is worth to emphasize that both these objects, Q_p and $\mathbf{S}$, depend ONLY on two parameters; the rotation pattern $\underline{\varepsilon}$ and the correlation coefficient ρ. It is known that these two assumptions are satisfied if the coverage of the pattern $p = 1$ or $p = 2$ for any cascade scheme and $p = 3$ for 2-2-2 scheme. It is not known if they are satisfied in general. However numerical

experiments allow to formulate a conjecture that this is really the case. In these experiments we considered many different rotation patterns. For each such a pattern we considered several values for $\rho \in (-1,1)$. Having the rotation pattern $\underline{\varepsilon}$ and the value of ρ chosen, we built respective polynomial Q_p and matrix $\mathbf{S}$. Numerically we looked for roots of Q_p. Often these roots were complex, but when they were real they were located outside of the interval $(-1,1)$ in all the experiments (that is, ASSUMPTION I was satisfied). Then we tried to solve numerically the equation $\mathbf{S}\underline{c} = (1,0,\ldots,0) \in \mathbb{R}^{ph+h+1}$. Again, in all the experiments we obtained the unique solution, meaning that $\mathbf{S}$ was of full rank (that is, ASSUMPTION II was also satisfied). We do believe that both the assumptions are always satisfied but a mathematical proof of these facts is probably hard, though a paper with the proof that ASSUMPTION I is satisfied for any cascade pattern with a single gap of any size and any $\rho \in (-1,1)$ is under preparation.

There is other type of limitations of the method we propose - they are due to the model constraints. In particular, in the model the correlations are exponential (as in the original Patterson model). This property is very important for the argument we use, e.g., it makes the covariance matrix $\mathbf{C}$ nilpotent of degree N, that is N is the smallest value of j such that $\mathbf{C}^j = 0$. Moreover, it has been observed (see Example 4.5 in Kowalski 2009) that other covariance models may lead to major difficulties in analysis of the formula for the variance of the estimators. There is a possibility that some reasonable departures from the exponential correlation assumption, as e.g., $\mathbb{C}\mathrm{ov}\left(X_{i,j}, X_{k,l}\right) = \theta + (1-\theta)\rho^{|j-l|}\delta_{i,k}$ for a $\theta \in [0,1]$ (see Lent, Miller, Cantwell and Duff (1999), in particular their Table 1, its discussion as well as additional references) can lead to treatable formulas for the variance. Such a covariance model is probably the first one to look at in any future research aiming at extension of the model.

In the model we also assumed that expectations on a given occasion are all the same and depend only on the occasion number: $\mathbb{E}X_{i,j} = \mu_j$. However other models may be of interest, e.g., $\mathbb{E}X_{i,j} = \mu_j + a_i$ (see Bailar 1975). Here the adjustments a_i can be understood as time-in-sample-bias caused by the number of occasions in which unit i participated in the survey. Of course, if a_i is known, there is no problem: just adjust $X_{i,j}$ by subtracting a_i and use the approach we developed. If it is not known, the operational (but not mathematical) solution would be to adjust $X_{i,j}$'s with suitable estimators of a_i's (obtained outside the model we analyze). The exact mathematical solution is not known and is worth to pursue.

Another aspect, which is of interest within the model considered in this paper, is the question of recurrence for the BLUE of a change of the mean $\mu_t - \mu_{t-1}$. We do believe that this question can be approached through the methods developed in this paper. Nevertheless, we expect it will need a lot of work in careful adaptations of the algebraic techniques used above.

It is worth also to mention that the model considered in the paper has an infinite time horizon, why there is always finite number of occasions in real surveys. As already mentioned in Introduction, the results we obtained seem to be reasonable approximation of the finite horizon case, when coefficients of recursion (1.2) depend on t. In particular, numerical experiments, performed for a wide range of $\rho \in (-1,1)$ and several different cascade patterns ε, show that e.g., the value of the coefficients $a_i^{(t)}$ (for the finite horizon) was roughly the same as a_i (for the infinite horizon) already for $t \approx 10$. The same behavior was observed for the variances of the estimators. Nevertheless, the convergence has been

mathematically established only in the case $p = 1$. Analytical bounds for the speed of convergence at present seem also to be out of reach.

It is interesting to know how the estimators, obtained here, work in real surveys. Such question needs access to real data and gaining some interest of practitioners in the theoretical solutions we proposed. Very likely the exact formulas given in Theorem 3.1 may need some adjustments due to the discussed limitations of the model.

6 Appendix

6.1 Algebra of shift operators

In the first part of Appendix we introduce and analyze an algebraic operator formalism which is crucial for the proof of our main result (given in Subsection 6.2).

For a sequence of vectors $\underline{\bar{x}} = (\underline{x}_0, \underline{x}_1, \underline{x}_2, \ldots), \underline{x}_i \in \mathbb{R}^N$, define shifts to the left and to the right by

$$\begin{aligned} \mathcal{L}(\underline{\bar{x}}) &= (\underline{x}_1, \underline{x}_2, \underline{x}_3, \ldots) \quad \text{left shift,} \\ \mathcal{R}(\underline{\bar{x}}) &= (\underline{0}, \underline{x}_0, \underline{x}_1, \ldots) \quad \text{right shift.} \end{aligned}$$

Note that $\mathcal{L}\mathcal{R} = \mathcal{I}$ (identity), but

$$(\mathcal{I} - \mathcal{R}\mathcal{L})\,\underline{\bar{x}} = (\underline{x}_0, \underline{0}, \underline{0}, \ldots) = \underline{x}_0 \bar{e}, \tag{6.1}$$

where $\bar{e} = (1, 0, 0, \ldots)$.

For any $M \times N$ matrix $\mathbf{A}$ define

$$\mathbf{A}\underline{\bar{x}} = (\mathbf{A}\underline{x}_0, \mathbf{A}\underline{x}_1, \mathbf{A}\underline{x}_2, \ldots).$$

In particular, for a complex (real) number a, taking $\mathbf{A} = a\mathbf{I}$ we have

$$a\underline{\bar{x}} = (a\underline{x}_0, a\underline{x}_1, a\underline{x}_2, \ldots).$$

Moreover, by the above definitions, for any $i, j \geq 0$

$$\mathcal{R}^i \mathcal{L}^j \mathbf{A}\underline{\bar{x}} = \mathbf{A}\mathcal{R}^i \mathcal{L}^j \underline{\bar{x}}.$$

For a constant sequence of vectors $\underline{\bar{x}} = (\underline{x}, \underline{x}, \underline{x}, \ldots)$ we have $\mathcal{L}\underline{\bar{x}} = \underline{\bar{x}}$ and thus for any $i, j \geq 0$

$$\mathcal{L}^i \mathcal{R}^j \underline{\bar{x}} = \begin{cases} \underline{\bar{x}}, & \text{for } i \geq j, \\ \mathcal{R}^{j-i} \underline{\bar{x}}, & \text{for } i < j. \end{cases} \tag{6.2}$$

If $N = 1$ we write $\underline{\bar{y}} = \bar{y} = (y_0, y_1, y_2 \ldots)$, $y_i \in \mathbb{R}$, and $\mathrm{L} := \mathcal{L}, \mathrm{R} := \mathcal{R}$. Note that, for $\bar{y} = \left(y^n\right)_{n \geq 0}$ we have

$$\mathrm{L}^j \overline{y} = y^j \overline{y} \tag{6.3}$$

and thus

$$\mathrm{L}^j \mathrm{R}^i \overline{y} = \begin{cases} y^{j-i}\overline{y}, & \text{for } j \geq i, \\ \mathrm{R}^{i-j}\overline{y}, & \text{for } j < i. \end{cases}$$

For any $\overline{y} = (y_n)_{n\geq 0}$ and any $\underline{\overline{x}} = (\underline{x}_n)_{n\geq 0}$ define $\overline{y}\underline{\overline{x}} = (y_n \underline{x}_n)_{n\geq 0}$. Then for any complex (real) numbers α, β, any $M \times N$ matrices $\mathbf{A}, \mathbf{B}$, any $i, j, k, m \geq 0$,

$$\left(\alpha \mathbf{A}\mathcal{R}^i\mathcal{L}^j + \beta \mathbf{B}\mathcal{L}^m\mathcal{R}^k\right)\overline{y}\underline{\overline{x}} = \left(\alpha \mathrm{R}^i\mathrm{L}^j\overline{y}\right)\left(\mathbf{A}\mathcal{R}^i\mathcal{L}^j\underline{\overline{x}}\right) + \left(\beta \mathrm{L}^m\mathrm{R}^k\overline{y}\right)\left(\mathbf{B}\mathcal{L}^m\mathcal{R}^k\underline{\overline{x}}\right). \tag{6.4}$$

Note also that if $\underline{\overline{x}} = (\underline{x}, \underline{x}, \ldots)$ is a constant sequence, then

$$\mathcal{R}^i\mathcal{L}^j\overline{y}\underline{\overline{x}} = \left(\mathrm{R}^i\mathrm{L}^j\overline{y}\right)\underline{\overline{x}} \quad \text{and} \quad \mathcal{L}^j\mathcal{R}^i\overline{y}\underline{\overline{x}} = \left(\mathrm{L}^j\mathrm{R}^i\overline{y}\right)\underline{\overline{x}}. \tag{6.5}$$

Lemma 6.1 *Let $v_i, i = 1, \ldots, p$, be functions defined in (3.11), where $a_1, \ldots, a_p$ are arbitrary numbers. Let $\underline{\overline{x}} = (\underline{x}, \underline{x}, \ldots)$ and $\overline{y} = \left(y^n\right)_{n\geq 0}$. Then for any $i = 1, \ldots, p$*

$$\mathcal{L}^i\left(\mathcal{I} - \sum_{j=1}^{p} a_j\mathcal{R}^j\right) = \left(\mathcal{L}^p - \sum_{j=1}^{p} a_j\mathcal{L}^{p-j}\right)\mathcal{R}^{p-i}, \tag{6.6}$$

$$(\mathcal{I} - \mathcal{R}\mathcal{L})\left(\mathcal{L}^p - \sum_{j=1}^{p} a_j\mathcal{L}^{p-j}\right)\mathcal{R}^{p-i}\overline{y}\underline{\overline{x}} = v_i(y)(\underline{x}_0, \underline{0}, \underline{0}, \ldots) \tag{6.7}$$

and

$$\left(\mathcal{L}^p - \sum_{j=1}^{p} a_j\mathcal{L}^{p-j}\right)\overline{y}\underline{\overline{x}} = v_p(y)\,\overline{y}\underline{\overline{x}}. \tag{6.8}$$

Proof. First, we prove (6.8). By (6.4)

$$\left(\mathcal{L}^p - \sum_{j=1}^{p} a_j\mathcal{L}^{p-j}\right)\overline{y}\underline{\overline{x}} = \left(\mathrm{L}^p\overline{y}\right)\mathcal{L}^p\underline{\overline{x}} - \sum_{j=1}^{p} a_j\left(\mathrm{L}^{p-j}\overline{y}\right)\left(\mathcal{L}^{p-j}\underline{\overline{x}}\right).$$

Note that $\mathrm{L}^k\overline{y} = y^k\overline{y}$ and $\mathcal{L}^k\underline{\overline{x}} = \underline{\overline{x}}$ for any $k = 0, 1, \ldots$. Therefore

$$\left(\mathcal{L}^p - \sum_{j=1}^{p} a_j\mathcal{L}^{p-j}\right)\overline{y}\underline{\overline{x}} = \left[\left(y^p - \sum_{m=1}^{p} a_m y^{p-m}\right)\overline{y}\right]\underline{\overline{x}}.$$

Now (6.8) follows by the definition (3.11) for $i = p$.

Again, from (6.2), (6.4) and (6.5) it follows that

$$(\mathcal{I} - \mathcal{R}\mathcal{L})\left(\mathcal{L}^p - \sum_{j=1}^{p} a_j\mathcal{L}^{p-j}\right)\mathcal{R}^{p-i}\overline{y}\underline{\overline{x}} = \left[(\mathrm{I} - \mathrm{RL})\left(\mathrm{L}^p - \sum_{j=1}^{p} a_j\mathrm{L}^{p-j}\right)\mathrm{R}^{p-i}\overline{y}\right]\underline{\overline{x}}.$$

Since for any $k \in \{0,1,\ldots,p\}$

$$\left(\mathrm{L}^p - \sum_{j=1}^{p} a_j \mathrm{L}^{p-j}\right)\mathrm{R}^{p-k}\overline{y} = y^k\overline{y} - \sum_{j=1}^{k} a_j y^{k-j}\overline{y} - \sum_{j=k+1}^{p} a_j \mathrm{R}^{j-k}\overline{y} = v_k(y)\,\overline{y} - \sum_{j=k+1}^{p} a_j \mathrm{R}^{j-k}\overline{y}$$

then

$$(\mathrm{I} - \mathrm{RL})\left(\mathrm{L}^p - \sum_{j=1}^{p} a_j \mathrm{L}^{p-j}\right)\mathrm{R}^{p-k}\overline{y} = v_k(y)\,\overline{e}$$

and thus (6.7) follows.

The identity (6.6) follows by (6.2) since

$$\mathcal{L}^i\left(\mathcal{I} - \sum_{j=1}^{p} a_j \mathcal{R}^j\right) = \mathcal{L}^i - \sum_{j=1}^{p} a_j \mathcal{L}^i\mathcal{R}^j = \mathcal{L}^p\mathcal{R}^{p-i} - \sum_{j=1}^{p} a_j \mathcal{L}^{p-j}\mathcal{R}^{p-i}.$$

Lemma 6.2 *Let $\mathcal{D}$ be an operator on the space of sequences of vectors from $\mathbb{R}^N$ defined by*

$$\mathcal{D} = \mathcal{I} + \sum_{k=1}^{N-1}\left(\mathbf{C}^k\mathcal{L}^k + \left(\mathbf{C}^T\right)^k\mathcal{R}^k\right), \tag{6.9}$$

where $\mathbf{C}$ is the covariance matrix defined in Section 2.

The operator $\mathcal{D}$ is invertible and

$$\mathcal{D}^{-1} = \left(\mathcal{I} - \mathbf{C}^T\mathcal{R}\right)\mathbf{\Delta}\left(\mathcal{I} - \mathbf{C}\mathcal{L}\right). \tag{6.10}$$

Proof. Note that $\mathbf{I} - \mathbf{C}\mathbf{C}^T = \mathrm{diag}\left(1-\rho^2,\ldots,1-\rho^2,1\right)$. Consequently, $\mathbf{\Delta} = \left(\mathbf{I} - \mathbf{C}\mathbf{C}^T\right)^{-1}$ is well defined. Note also that $\sum_{k=0}^{N-1}\mathbf{C}^k\mathcal{L}^k$ is invertible and its inverse is $\mathcal{I} - \mathbf{C}\mathcal{L}$. Similarly, $\sum_{k=0}^{N-1}(\mathbf{C}^T)^k\mathcal{R}^k$ is invertible and its inverse is $\mathcal{I} - \mathbf{C}^T\mathcal{R}$.

Therefore

$$\begin{aligned}
\left[\left(\mathcal{I} - \mathbf{C}^T\mathcal{R}\right)\mathbf{\Delta}\left(\mathcal{I} - \mathbf{C}\mathcal{L}\right)\right]^{-1} &= \left(\mathcal{I} - \mathbf{C}\mathcal{L}\right)^{-1}\mathbf{\Delta}^{-1}\left(\mathcal{I} - \mathbf{C}^T\mathcal{R}\right)^{-1} \\
&= \left(\sum_{k=0}^{N-1}\mathbf{C}^k\mathcal{L}^k\right)\left(\mathbf{I} - \mathbf{C}\mathbf{C}^T\right)\left(\sum_{j=0}^{N-1}\left(\mathbf{C}^T\right)^j\mathcal{R}^j\right) \\
&= \sum_{k,j=0}^{N-1}\mathbf{C}^k\left(\mathbf{C}^T\right)^j\mathcal{L}^k\mathcal{R}^j - \sum_{k,j=1}^{N-1}\mathbf{C}^k\left(\mathbf{C}^T\right)^j\mathcal{L}^k\mathcal{R}^j \\
&= \mathcal{D} + \sum_{k,j=1}^{N-1}\mathbf{C}^k\left(\mathbf{C}^T\right)^j\mathcal{L}^{k-1}\left(\mathcal{L}\mathcal{R} - \mathcal{I}\right)\mathcal{R}^{j-1} \\
&= \mathcal{D}.
\end{aligned}$$

6.2 Proof of the recurrence

Proof of Theorem 3.1. Note first that since $d_1, \ldots, d_p$ are either real or come in conjugate pairs (see Remark 3.1) it follows from (3.10) that $a_1, \ldots, a_p$ are real numbers.

Recall that $\underline{e}_0 = \underline{1}$ and denote $\overline{\underline{e}}_j = (\underline{e}_j, \underline{e}_j, \ldots), j \in H' = \{0\} \bigcup H$. Recall that the $N \times N$ diagonal matrix $\mathbf{\Delta}$ is defined as

$$\mathbf{\Delta} = \left(\mathbf{I} - \mathbf{C}\mathbf{C}^T\right)^{-1} = \frac{1}{1-\rho^2}\,\mathrm{diag}\left(1, \ldots, 1, 1-\rho^2\right).$$

With $d_1, \ldots, d_p$ and $\underline{c}$ as defined in Theorem 3.1 let (see (6.10))

$$\overline{\underline{w}} = (\underline{w}_0, \underline{w}_1, \ldots) = \mathcal{D}^{-1} \sum_{m=1}^{p} \sum_{j \in H'} c_{j,m}\, \overline{d}_m \overline{\underline{e}}_j, \tag{6.11}$$

where $\overline{d}_m = \left(1, d_m, d_m^2, \ldots\right), m = 1, \ldots, p$. Note that $\|\underline{w}_i\|$ (the length of the vector $\underline{w}_i$) is of order $\left(\max_{1 \le m \le p} |d_m|\right)^i, i = 0, 1, \ldots$. By Remark 3.1 and ASSUMPTION II we have $\max_{1 \le m \le p} |d_m| \in (0,1)$. Hence (2.1) is a correct definition of a random series (with bounded variance).

Consequently, it suffices to show that:

1. The sequence $\overline{\underline{w}}$ defined in (6.11) is the sequence of optimal weights. To this end we note that the variance of any linear estimator $\sum_{i=0}^{\infty} \underline{u}_i^T \underline{X}_i, \underline{u}_i \in \mathbb{R}^N, i = 0, 1, \ldots,$ has the form

$$\mathbb{V}\mathrm{ar} \sum_{i=0}^{\infty} \underline{u}_i^T \underline{X}_i = \sum_{i=0}^{\infty} \underline{u}_i^T \underline{u}_i + 2 \sum_{i=0}^{\infty} \sum_{k=1}^{N-1} \underline{u}_i^T \mathbf{C}^k \underline{u}_{i+k}. \tag{6.12}$$

   We need to show that $\overline{\underline{u}} = (\underline{u}_i)_{i \ge 0} := \overline{\underline{w}}$ with $\overline{\underline{w}}$ as defined in (6.11) minimize this expression under the constraints (2.2) and (2.3). Since the above variance as a function of $\overline{\underline{u}}$ is convex then the problem has the unique solution. Using the standard Lagrange method, that is differentiating the Lagrange function (with multipliers $(\lambda_{j,i})_{j \in H', i \ge 0}$)

$$V(\overline{\underline{u}}) = \sum_{i=0}^{\infty} \underline{u}_i^T \underline{u}_i + 2 \sum_{i=0}^{\infty} \sum_{k=1}^{N-1} \underline{u}_i^T \mathbf{C}^k \underline{u}_{i+k} - 2 \sum_{i=0}^{\infty} \sum_{j \in H'} \lambda_{j,i} \underline{u}_i^t \underline{e}_j,$$

   with respect to $(\underline{u}_i)_{i \ge 0}$ and comparing the derivatives to zero, equivalently, we need to show that there exist real numbers (Lagrange multipliers) $\lambda_{j,l}, j \in H', l = 0, 1, \ldots,$ such that

$$\mathcal{D}\overline{\underline{w}} = \left[\mathcal{I} + \sum_{k=1}^{N-1} \left(\mathbf{C}^k \mathcal{L}^k + \left(\mathbf{C}^T\right)^k \mathcal{R}^k\right)\right] \overline{\underline{w}} = \overline{\underline{\Lambda}}, \tag{6.13}$$

   where $\overline{\underline{w}}$ is defined in (6.11) and $\overline{\underline{\Lambda}} = (\underline{\Lambda}_0, \underline{\Lambda}_1, \ldots)$ with

$$\underline{\Lambda}_l = \sum_{j \in H'} \lambda_{j,l} \underline{e}_j, \quad l = 0, 1, \ldots$$

2. The constraints (2.2) and (2.3) are satisfied for $\underline{\overline{w}}$ as defined in (6.11).
3. The basic recurrence (3.9) holds true with $\underline{\overline{w}}$ defined in (6.11), that is the sequence $\underline{\overline{r}}$ defined by

$$\underline{\overline{r}} := \left(\mathcal{I} - \sum_{m=1}^{p} a_m \mathcal{R}^m \right) \underline{\overline{w}} \tag{6.14}$$

has to satisfy

$$\mathcal{L}^{p+1} \underline{\overline{r}} = \underline{\overline{0}} \tag{6.15}$$

and for any $i = 0,1,\ldots,p$

$$(\mathcal{I} - \mathcal{R}\mathcal{L})\,\mathcal{L}^i \underline{\overline{r}} = \sum_{m=1}^{p} \left[\left(v_i(d_m)\mathbf{I} - v_{i-1}(d_m)\mathbf{C}^T \right) \mathbf{N}(d_m) \sum_{j \in H'} c_{j,m} \underline{e}_j \right] \overline{e}, \tag{6.16}$$

where $\mathbf{N}(d) = \mathbf{\Delta}(\mathbf{I} - d\mathbf{C})$.

Ad. 1. We will show that (6.13) holds with

$$\lambda_{j,l} = \sum_{m=1}^{p} c_{j,m} d_m^l, \qquad j \in H', l = 0,1,\ldots \tag{6.17}$$

By definition (6.11) of $\underline{\overline{w}}$ we have

$$\mathcal{D}\underline{\overline{w}} = \sum_{m=1}^{p} \sum_{j \in H'} c_{j,m} \overline{d}_m \underline{\overline{e}}_j = \left(\sum_{j \in H'} \sum_{m=1}^{p} c_{j,m} d_m^l \underline{e}_j, l = 0,1,\ldots \right).$$

Therefore, by definition of $\lambda_{j,l}$'s we obtain

$$\mathcal{D}\underline{\overline{w}} = \left(\sum_{j \in H'} \lambda_{j,l} \underline{e}_j \right) = (\underline{\Lambda}_0, \underline{\Lambda}_1, \ldots) = \underline{\overline{\Lambda}}.$$

To see that $\lambda_{j,l}$ as defined through (6.17) are real numbers take first conjugates of both sides of $\mathbf{S}\underline{c} = \underline{e}$. Note that

$$\mathbf{S}^* = \mathbf{S}^*(d_1,\ldots,d_p) = \mathbf{S}(d_1^*,\ldots,d_p^*).$$

Since $d_1,\ldots,d_p$ are either real or come in conjugate pairs (see Rem. 3.1) the equation $\mathbf{S}^*\underline{c}^* = \underline{e}$ implies that for any $j \in H'$ and any $m = 1,\ldots,p$ either $\Im d_m = 0$ and then $c_{j,m}$ is real or $\Im d_m \neq 0$ and then there exists $n \neq m$ (with $d_n^* = d_m$) such that $c_{j,n}^* = c_{j,m}$. Therefore the quantities $c_{j,m} d_m^l$ in (6.17) are either real or come in conjugate pairs. Consequently, by (6.17) it follows that $\lambda_{j,l}$ is real.

Ad. 2. Note that applying (6.1) and (6.4) to (6.11) after an easy algebra we get

$$\underline{w}_0 = \sum_{m=1}^{p} \sum_{j \in H'} c_{j,m} \mathbf{N}(d_m) \underline{e}_j$$

and

$$\underline{w}_i = \sum_{m=1}^{p} \sum_{j \in H'} c_{j,m} d_m^{i-1} \left(d_m \mathbf{I} - \mathbf{C}^T \right) \mathbf{N}(d_m) \underline{e}_j, \quad i = 1, 2, \ldots$$

Let us rewrite the constraints (2.2) and (2.3) using the above formulas for $\underline{w}_0$ and $\underline{w}_i, i \geq 1$. The constraint (2.2) for $i = 0$ with $\underline{w}_0$ as defined above takes on the form

$$\sum_{m=1}^{p} \sum_{j \in H'} c_{j,m} \underline{1}^T \mathbf{N}(d_m) \underline{e}_j = 1 \tag{6.18}$$

and for $i \geq 1$

$$\sum_{m=1}^{p} \sum_{j \in H'} c_{j,m} d_m^{i-1} \underline{1}^T \left(d_m \mathbf{I} - \mathbf{C}^T \right) \mathbf{N}(d_m) \underline{e}_j = 0. \tag{6.19}$$

The constraint (2.3) for $i = 0$, that is for $\underline{w}_0$, has the form

$$\sum_{m=1}^{p} \sum_{j \in H'} c_{j,m} \underline{e}_k^T \mathbf{N}(d_m) \underline{e}_j = 0, \quad k \in H. \tag{6.20}$$

For $i > 0$ it has the form

$$\sum_{m=1}^{p} \sum_{j \in H'} c_{j,m} d_m^{i-1} \underline{e}_k^T \left(d_m \mathbf{I} - \mathbf{C}^T \right) \mathbf{N}(d_m) \underline{e}_j = 0, \quad k \in H. \tag{6.21}$$

Note that $N \times N$ matrix

$$\mathbf{N}(d) = \frac{1}{1-\rho^2} \begin{bmatrix} 1 & -\rho d & \ddots & 0 & 0 \\ 0 & 1 & \ddots & \ddots & 0 \\ \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & \ddots & \ddots & 1 & -\rho d \\ 0 & 0 & \ddots & 0 & 1-\rho^2 \end{bmatrix}$$

and $\left(d\mathbf{I} - \mathbf{C}^T \right) \mathbf{N}(d) = \frac{d}{1-\rho^2} \mathbf{H}_N(d)$ - see (3.8). Thus, by elementary computations, we get

$$\underline{e}_k^T \mathbf{N}(d) \underline{e}_j = \frac{1}{1-\rho^2} \begin{cases} (N-1)(1-d\rho) + 1 - \rho^2, & k = j = 0, \\ 1 - d\rho, & k = 0, j \in H \text{ or } k \in H, j = 0, \\ 1, & k = j, \\ -d\rho, & k = j-1, \\ 0, & \text{otherwise}, \end{cases} \quad k, j \in H \tag{6.22}$$

and

$$\underline{e}_k^T \left(d\mathbf{I} - \mathbf{C}^T\right) \mathbf{N}(d)\, \underline{e}_j$$

$$= \frac{1}{1-\rho^2} \begin{cases} (N-1)(1-d\rho)(d-\rho) + d\left(1-\rho^2\right), & k = j = 0, \\ (1-d\rho)(d-\rho), & k = 0,\ j \in H \text{ or } k \in H,\ j = 0, \\ -\rho, & k = j+1, \\ d\left(1+\rho^2\right), & k = j, \\ -d^2\rho, & k = j-1, \\ 0, & \text{otherwise}, \end{cases} \quad k, j \in H. \tag{6.23}$$

Due to (6.22) and (6.23), the constraints (6.18), (6.19), (6.20) and (6.21) can be rewritten in a matrix form as

$$\begin{bmatrix} \widetilde{\mathbf{G}}(d_1) & \widetilde{\mathbf{G}}(d_2) & \cdots & \widetilde{\mathbf{G}}(d_p) \\ \bar{\mathbf{G}}(d_1) & \bar{\mathbf{G}}(d_2) & \cdots & \bar{\mathbf{G}}(d_p) \\ d_1\bar{\mathbf{G}}(d_1) & d_2\bar{\mathbf{G}}(d_2) & \cdots & d_p\bar{\mathbf{G}}(d_p) \\ \vdots & \vdots & \ddots & \vdots \\ d_1^i\bar{\mathbf{G}}(d_1) & d_2^i\bar{\mathbf{G}}(d_2) & \cdots & d_p^i\bar{\mathbf{G}}(d_p) \\ \vdots & \vdots & \ddots & \vdots \end{bmatrix} \underline{c} = \bar{e}, \tag{6.24}$$

where $\widetilde{\mathbf{G}}(d)$ is defined through (3.5) and (3.6),

$$\bar{\mathbf{G}}(d) = \frac{d}{1-\rho^2} \begin{bmatrix} \mathbf{H}_{11}(d) & \mathbf{H}_{12}(d) \\ \mathbf{H}_{21}(d) & \mathbf{H}_{22}(d) \end{bmatrix}$$

with

$$\mathbf{H}_{11}(d) = (N-1)(1-\rho d)(1-\rho/d) + 1 - \rho^2,$$

$$\mathbf{H}_{12} = \mathbf{H}_{21}^T = (1-\rho d)(1-\rho/d)\, \underline{1}_h^T,$$

$$\mathbf{H}_{22}(d) = \operatorname{diag}\left(\mathbf{H}_1(d), \ldots, \mathbf{H}_s(d)\right),$$

and matrices $\mathbf{H}_i(d), i = 1, \ldots, s,$ are defined in (3.8).

The infinite matrix at the left hand side of (6.24) can be written as

$$\begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{I} & \cdots & \mathbf{I} \\ \mathbf{0} & d_1\mathbf{I} & d_2\mathbf{I} & \cdots & d_p\mathbf{I} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & d_1^i\mathbf{I} & d_2^i\mathbf{I} & \cdots & d_p^i\mathbf{I} \\ \vdots & \vdots & \vdots & \ddots & \vdots \end{bmatrix} \begin{bmatrix} \widetilde{\mathbf{G}}(d_1) & \widetilde{\mathbf{G}}(d_2) & \cdots & \widetilde{\mathbf{G}}(d_p) \\ \bar{\mathbf{G}}(d_1) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \bar{\mathbf{G}}(d_2) & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \bar{\mathbf{G}}(d_p) \end{bmatrix},$$

where $\mathbf{I} = \mathbf{I}_{h+1}$ and $\mathbf{0} = \mathbf{0}_{h+1}$ are, respectively, $(h+1) \times (h+1)$ unit and zero matrices. Note that the first matrix in the product above is of full rank and can be written as

$$\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 1 & \cdots & 1 \\ 0 & d_1 & d_2 & \cdots & d_p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & d_1^i & d_2^i & \cdots & d_p^i \\ \vdots & \vdots & \vdots & \ddots & \vdots \end{bmatrix} \otimes \mathbf{I}_{h+1}.$$

Therefore (6.24) is equivalent to

$$\begin{bmatrix} \widetilde{\mathbf{G}}(d_1) & \widetilde{\mathbf{G}}(d_2) & \cdots & \widetilde{\mathbf{G}}(d_p) \\ \bar{\mathbf{G}}(d_1) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \bar{\mathbf{G}}(d_2) & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \bar{\mathbf{G}}(d_p) \end{bmatrix} \underline{c} = (1, 0, \ldots, 0)^T \in \mathbb{R}^{(p+1)(h+1)}. \tag{6.25}$$

Assume that we prove that $(h+1) \times (h+1)$ matrices $\bar{\mathbf{G}}(d_m), m = 1, \ldots, p,$ are singular. Note that $d[\mathbf{H}_{21}(d), \mathbf{H}_{22}(d)] = \mathbf{G}(d)$ due to (3.7). Therefore, the definition (3.4) of $\mathbf{S}$ implies that (6.25) is equivalent to $\mathbf{S}\underline{c} = (1, 0, \ldots, 0) \in \mathbb{R}^{ph+h+1}$. It is obtained from (6.25) by deleting all rows determined through first rows of matrices $\bar{\mathbf{G}}(d_m), m = 1, \ldots, p$. And the equation $\mathbf{S}\underline{c} = (1, 0, \ldots, 0)$ follows by ASSUMPTION II and the definition of $\underline{c}$.

Consequently, it suffices to show that $\det \bar{\mathbf{G}}(d_m) = 0, m = 1, \ldots, p$. That is, we need to check that

$$0 = \det \begin{bmatrix} \mathbf{H}_{11}(d_m) & \mathbf{H}_{12}(d_m) \\ \mathbf{H}_{21}(d_m) & \mathbf{H}_{22}(d_m) \end{bmatrix}$$

for any $m = 1, \ldots, p$.

Note that with $d = d_m$ the right hand side can be written as

$$\det \mathbf{H}_{22}(d) \det \left[\mathbf{H}_{11}(d) - \mathbf{H}_{12}(d) \mathbf{H}_{22}^{-1}(d) \mathbf{H}_{21}(d) \right]$$

and

$$\det \mathbf{H}_{22}(d) = \prod_{i=1}^{s} \det \mathbf{H}_{m_i}(d). \tag{6.26}$$

Since $\mathbf{H}_m(d)$ can be decomposed as

$$\mathbf{H}_m(d) = \mathbf{D}_m^{-1} \mathbf{R}_m \mathbf{D}_m, \tag{6.27}$$

where $\mathbf{D}_m = \text{diag}\left(1, d, d^2, \ldots, d^{m-1}\right)$ and $\mathbf{R}_m$ is defined in (3.2) we see that

$$\det \mathbf{H}_m(d) = 1 + \rho^2 + \ldots + \rho^{2m} \neq 0.$$

Now, from (6.26) it follows that $\det \mathbf{H}_{22} \neq 0$.

On the other hand

$$\det\left[\mathbf{H}_{11}(d) - \mathbf{H}_{12}(d)\mathbf{H}_{22}^{-1}(d)\mathbf{H}_{21}(d)\right] = (N-1)\alpha(\rho, d) + 1 - \rho^2 - \alpha^2(\rho, d)\sum_{j=1}^{s} \underline{1}^T \mathbf{H}_{m_j}^{-1} \underline{1}, \quad (6.28)$$

where $\alpha(\rho, d) = 1 + \rho^2 - \left(d + d^{-1}\right)\rho$.

The decomposition (6.27) of $\mathbf{H}_m$ gives

$$\underline{1}^T \mathbf{H}_m^{-1} \underline{1} = \text{tr}\left(\underline{1}^T \mathbf{D}_m^{-1} \mathbf{R}_m^{-1} \mathbf{D}_m \underline{1}\right) = \text{tr}\left(\mathbf{D}_m \underline{1}\underline{1}^T \mathbf{D}_m^{-1} \mathbf{R}_m^{-1}\right).$$

Moreover, since $\text{tr}(\mathbf{A}) = \text{tr}\left(\mathbf{A}^T\right)$

$$\underline{1}^T \mathbf{H}_m^{-1} \underline{1} = \text{tr}\left(\left(\mathbf{D}_m \underline{1}\underline{1}^T \mathbf{D}_m^{-1} \mathbf{R}_m^{-1}\right)^T\right) = \text{tr}\left(\mathbf{R}_m^{-1} \mathbf{D}_m^{-1} \underline{1}\underline{1}^T \mathbf{D}_m\right) = \text{tr}\left(\mathbf{D}_m^{-1} \underline{1}\underline{1}^T \mathbf{D}_m \mathbf{R}_m^{-1}\right).$$

Combining the last two expressions for $\underline{1}^T \mathbf{H}_m^{-1} \underline{1}$ we get

$$\underline{1}^T \mathbf{H}_m^{-1} \underline{1} = \text{tr}\left(\tfrac{1}{2}\left(\mathbf{D}_m \underline{1}\underline{1}^T \mathbf{D}_m^{-1} + \mathbf{D}_m^{-1} \underline{1}\underline{1}^T \mathbf{D}_m\right)\mathbf{R}_m^{-1}\right).$$

Note that

$$\left(\mathbf{D}_m \underline{1}\underline{1}^T \mathbf{D}_m^{-1} + \mathbf{D}_m^{-1} \underline{1}\underline{1}^T \mathbf{D}_m\right)_{ij} = d^{|i-j|} + d^{-|i-j|},$$

and that

$$\tfrac{1}{2}\left(d^k + d^{-k}\right) = T_k\left(\tfrac{1}{2}\left(d + d^{-1}\right)\right), \quad k = 0, 1, \ldots,$$

where (T_k) is the k^{th} Chebyshev polynomials of the first type.

Thus

$$\underline{1}^T \mathbf{H}_m^{-1} \underline{1} = \text{tr}\mathbf{T}_m(x)\mathbf{R}_m^{-1},$$

where $x = x(d) = \tfrac{1}{2}\left(d + d^{-1}\right)$ and the matrix $\mathbf{T}_m$ is defined in (3.1). Plugging this expression to (6.28) we find out that

$$\det\left(\mathbf{H}_{11}(d) - \mathbf{H}_{12}(d)\mathbf{H}_{22}^{-1}(d)\mathbf{H}_{21}(d)\right) = Q_p(x(d)),$$

where Q_p is the polynomial defined in (3.3). By ASSUMPTION I $Q_p(x(d_m)) = 0$, thus the above equality gives $\det \bar{\mathbf{G}}(d_m) = 0, m = 1, \ldots, p$. Finally, we conclude that the constraints (2.2) and (2.3) are satisfied and thus the proof of point 2 is completed.

Ad. 3. First, we will show that for $\underline{\overline{r}}$ defined by (6.14) the identity (6.15) holds. To this end observe that by (6.6) for $i = p$, (6.10) and (6.13)

$$\mathcal{L}^{p+1}\left(\mathcal{I} - \sum_{m=1}^{p} a_m \mathcal{R}^m\right)\underline{\overline{w}} = \mathcal{L}\left(\mathcal{L}^p - \sum_{m=1}^{p} a_m \mathcal{L}^{p-m}\right)\mathcal{D}^{-1}\underline{\overline{\Lambda}} = \mathcal{L}\mathcal{D}^{-1}\left(\mathcal{L}^p - \sum_{m=1}^{p} a_m \mathcal{L}^{p-m}\right)\underline{\overline{\Lambda}}.$$

Note also that for any $j = 1, \ldots, p$ by (6.8)

$$\left(\mathcal{L}^p - \sum_{m=1}^{p} a_m \mathcal{L}^{p-m}\right)\overline{d}_j = v_p\left(d_j\right)\overline{d}_j.$$

By the definition (3.10) of $a_m, m = 1, \ldots, p$ it follows that $v_p\left(d_j\right) = 0$. Due to the definition of Λ through (6.17) we conclude that $\mathcal{L}^{p+1}\underline{\overline{r}} = \underline{\overline{0}}$.

In order to check (6.16) first we note that due to (6.10) it follows from (6.3) and (6.5) that for $\overline{y} = \left(y^n\right)_{n\geq 0}$ and $\underline{\overline{x}} = (\underline{x}, \underline{x}, \ldots)$

$$\mathcal{D}^{-1}\overline{y}\underline{\overline{x}} = \left(\mathcal{I} - \mathbf{C}^T\mathcal{R}\right)\mathbf{N}(y)\,\overline{y}\underline{\overline{x}}.$$

Therefore for any $i \geq 0$ any d_j and $\underline{e}_{j_k}$ by (6.6)

$$\begin{aligned}\mathcal{L}^i\left(\mathcal{I} - \sum_{m=1}^{p} a_m \mathcal{R}^m\right)\mathcal{D}^{-1}\overline{d}_j\underline{\overline{e}}_{j_k} &= \left(\mathcal{L}^p - \sum_{m=1}^{p} a_m \mathcal{L}^{p-m}\right)\mathcal{R}^{p-i}\overline{d}_j\mathbf{N}\left(d_j\right)\underline{\overline{e}}_{j_k} \\ &- \left(\mathcal{L}^p - \sum_{m=1}^{p} a_m \mathcal{L}^{p-m}\right)\mathcal{R}^{p-(i-1)}\overline{d}_j\mathbf{C}^T\mathbf{N}\left(d_j\right)\underline{\overline{e}}_{j_k}.\end{aligned}$$

Finally, we use (6.7) with $\overline{y} = \overline{d}_j, \underline{\overline{x}} = \mathbf{N}\left(d_j\right)\underline{\overline{e}}_{j_k}$ to the first part and with $\overline{y} = \overline{d}_j$, $\underline{\overline{x}} = \mathbf{C}^T\mathbf{N}\left(d_j\right)\underline{\overline{e}}_{j_k}$ to the second part of the expression at the right hand side of the equation above arriving at

$$(\mathcal{I} - \mathcal{R}\mathcal{L})\,\mathcal{L}^i\left(\mathcal{I} - \sum_{m=1}^{p} a_m \mathcal{R}^m\right)\mathcal{D}^{-1}\overline{d}_j\underline{\overline{e}}_{j_k} = \left(v_i\left(d_j\right)\mathbf{I} - v_{i-1}\left(d_j\right)\mathbf{C}^T\right)\mathbf{N}\left(d_j\right)\left(\underline{e}_{j_k}, \underline{0}, \underline{0}, \ldots\right).$$

Thus (6.16) holds true.

Finally we will prove the formula (3.12) for the variance of the BLUE $\hat{\mu}_t$. To this end we observe first that

$$\mathbb{C}\mathrm{ov}\left(\hat{\mu}_t, \underline{X}_{t-i}\right) = \underline{w}_i + \sum_{k=1}^{N-1}\mathbf{C}^k\underline{w}_{i+k} + \sum_{k=1}^{i\wedge(N-1)}\left(\mathbf{C}^T\right)^k\underline{w}_{i-k}$$

for any $i = 0, 1, \ldots$. On the other hand, due to (6.13), we see that the right hand side of the above equality is equal to $\underline{\Lambda}_i$. That is, for any $i = 0, 1, \ldots$

$$\mathbb{C}\mathrm{ov}\left(\hat{\mu}_t, \underline{X}_{t-i}\right) = \sum_{j\in H'}\lambda_{j,i}\underline{e}_j.$$

Now, we write

$$\mathbb{V}\mathrm{ar}\hat{\mu}_t = \sum_{i=0}^{\infty} \underline{w}_i^T \mathbb{C}\mathrm{ov}\left(\hat{\mu}_t, \underline{X}_{t-i}\right) = \sum_{i=0}^{\infty} \sum_{j \in H'} \lambda_{j,i} \underline{w}_i^T \underline{e}_j .$$

Due to the constraints (2.2) and (2.3) it follows from the above formula that $\mathbb{V}\mathrm{ar}\hat{\mu}_t = \lambda_{0,0}$. Thus, (3.12) follows from (6.17).

References

Australian Bureau of Statistics (2002). Labour Force Survey sample design. *Information Paper, Catalogue no. 6269.0*.

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

Bell, P. (2001). Comparison of alternative Labour Force Survey estimators. *Survey Methodology,* 27, 1, 53-63.

Binder, D.A., and Hidiroglou, M.A. (1988). Sampling in time. *Handbook of Statistics,* 6, 187-211.

Cantwell, P.J. (1988). Variance formulae for the generalized composite estimator under balanced one-level rotation plan. *SRD Research Report Census/SRD/88/26*, Bureau of the Census, Statistical Research Division, 1-16.

Cantwell, P.J. (1990). Variance formulae for composite estimators in rotation designs. *Survey Methodology,* 16, 1, 153-163.

Cantwell, P.J., and Caldwell, C.V. (1998). Examining the revisions in monthly retail and wholesale trade surveys under a rotation panel design. *Journal of Official Statistics,* 14, 47-54.

Ciepiela, P., Gniado, M., Wesołowski, J. and Wojtyś, M. (2012). Dynamic K-composite estimator for an arbitrary rotation scheme. *Statistics in Transition,* 13(1), 7-20.

Fuller, W.A. and Rao, J.N.K. (2001). A regression composite estimator with application to the Canadian Labour Force Survey. *Survey Methodology,* 27, 1, 45-51.

Gurney, M., and Daly, J.F. (1965). A multivariate approach to estimation in periodic sample surveys. *Proceedings of the Survey Research Methods Section,* American Statistical Association, 242-257.

Hansen, M.H., Hurwitz, W.N., Nisselson, H. and Steinberg, J. (1955). The redesign of the census current population survey. *Journal of the American Statistical Association,* 50, 701-719.

Kowalski, J. (2009). Optimal estimation in rotation patterns. *Journal of Statistical Planning and Inference,* 139, 1405-1420.

Lent, J., Miller, S.M., Cantwell, P.J. and Duff, M. (1999). Effects of composite weights on some estimates from current population survey. *Journal of Official Statistics,* 15(3), 431-448.

Lind, J.T. (2005). Repeated surveys and the Kalman filter. *The Econometrics Journal,* 9, 1-10.

McLaren, C.H., and Steel, D.G. (2000). The impact of different rotation patterns on the sampling variance of seasonally adjusted and trend estimates. *Survey Methodology,* 26, 2, 163-172.

Patterson, H.D. (1950). Sampling on successive occasions. *Journal of the Royal Statistical Society, Series B,* 12, 241-255.

Popiński, W. (2006). Development of the Polish Labour Force Survey. *Statistics in Transition,* 7(5), 1009-1030.

Rao, J.N.K., and Graham, J.E. (1964). Rotation designs for sampling on repeated occasions. *Journal of the American Statistical Association,* 59, 492-509.

Singh, M.P., Drew, J.D., Gambino, J.G. and Mayda, F. (1990). Methodology of the Canadian Labour Force Survey 1984-1990. Statistics Canada, Catalogue no. 71-526.

Singh, A.C., Kennedy, B. and Wu, S. (2001). Regression composite estimation for the Canadian Labour Force Survey with a rotating panel design. *Survey Methodology,* 27, 1, 33-44.

Steel, D., and McLaren, C. (2002). In search of a good rotation pattern. *Advances in Statistics, Combinatorics and Related Areas,* Singapore, World Scientific, 309-319.

Steel, D., and McLaren, C. (2008). Design and analysis of repeated surveys. Centre for Statist. Survey Meth., Univ. Wollonong, Working Paper 11-08 (2008), 1-13, http://ro.uow.edu.au/cssmwp/10.

Szarkowski, A., and Witkowski, J. (1994). The Polish Labour Force Survey. *Statistics in Transition,* 1(4), 467-483.

Towhidi, M., and Namazi-Rad, M.-R. (2010). An optimal method of estimation in rotation sampling. *Advanced Applied Statistics,* 15(2), 115-136.

U.S. Bureau of Census (2002). The Current Population Survey - Design and Methodology. Department of Commerce, Technical Paper 63.

Wesołowski, J. (2010). Recursive optimal estimation in Szarkowski rotation scheme. *Statistics in Transition,* 11(2), 267-285.

Yansaneh, I.S., and Fuller, W.A. (1998). Optimal recursive estimation for repeated surveys. *Survey Methodology,* 24, 1, 31-40.

Survey Methodology, June 2015
Vol. 41, No. 1, pp. 127-144
Statistics Canada, Catalogue No. 12-001-X

Optimal adjustments for inconsistency in imputed data

Jeroen Pannekoek and Li-Chun Zhang[1]

Abstract

Imputed micro data often contain conflicting information. The situation may e.g., arise from partial imputation, where one part of the imputed record consists of the observed values of the original record and the other the imputed values. Edit-rules that involve variables from both parts of the record will often be violated. Or, inconsistency may be caused by adjustment for errors in the observed data, also referred to as imputation in Editing. Under the assumption that the remaining inconsistency is not due to systematic errors, we propose to make adjustments to the micro data such that all constraints are simultaneously satisfied and the adjustments are minimal according to a chosen distance metric. Different approaches to the distance metric are considered, as well as several extensions of the basic situation, including the treatment of categorical data, unit imputation and macro-level benchmarking. The properties and interpretations of the proposed methods are illustrated using business-economic data.

Key Words: Edit-rules; Consistent micro-data; Optimization; Benchmarking.

1 Introduction

We are concerned with the task of reconciling conflicting information in imputed micro data. To illustrate, consider a small part of a record from a structural business survey given in Table 1.1. Two response patterns are postulated; one with only Turnover observed and one where also Employees and Wages are observed. There are many ways to impute the missing values in such a *recipient* record and the proposed adjustment methods apply irrespective of the imputation method used. The use of partial donor imputation is shown in Table 1.1, where the *donor* record is the 'nearest neighbour' from the same category of economic activity and closest to the recipient record with respect to Turnover for response pattern (I) and Employees, Turnover and Wages for response pattern (II). The imputation is said to be partial because a value of the donor is transferred to the receptor if and only if the corresponding one is missing in the recipient record.

Business records generally have to adhere to a number of accounting and logical constraints. For checking of the validity of a record these are referred to as edit-rules. For the example record here, suppose the following three edit-rules are formulated:

$$a1: \quad x_1 - x_5 + x_8 = 0 \quad (\text{Profit} = \text{Turnover} - \text{Total Costs})$$

$$a2: \quad x_5 - x_3 - x_4 = 0 \quad (\text{Turnover} = \text{Turnover main} + \text{Turnover other})$$

$$a3: \quad x_8 - x_6 - x_7 = 0 \quad (\text{Total Costs} = \text{Wages} + \text{Other costs}).$$

Partial donor imputation leads to violation of these edit-rules, which we refer to as the (*micro-level*) *consistency problem*: for response pattern (I), the first two edit-rules involving Turnover are violated; for response pattern (II), all three edit-rules are violated. To obtain a consistent record, *some* of the eight values (i.e., including both the observed and imputed ones) have to be changed. Now, in the two cases

1. Jeroen Pannekoek, Statistics Netherlands, Henri Faasdreef 312, 2492 JP Den Haag, The Netherlands. E-mail: j.pannekoek@cbs.nl; Li-Chun Zhang, University of Southampton, Social Statistics and Demography, Highfield SO17 1BJ, Southampton, UK and Statistics Norway, Kongensgate 6, Pb 8131 Dep, 0033 Oslo, Norway. E-mail: L.Zhang@soton.ac.uk.

here, it is possible to change only the imputed values to satisfy all the edit-rules, so let us consider adjustments of the imputed values for the moment.

Table 1.1
Data, missing data and donor values for variables in a business record. Employees (Number of employees); Turnover main (Turnover main activity); Turnover other (Turnover other activities); Turnover (Total turnover); Wages (Costs of wages and salaries)

Variable	Name	Response (I)	Response (II)	Donor Values
x_1	Profit			330
x_2	Employees		25	20
x_3	Turnover Main			1,000
x_4	Turnover Other			30
x_5	Turnover	950	950	1,030
x_6	Wages		550	500
x_7	Other Costs			200
x_8	Total Costs			700

Traditional adjustment methods, such as the prorating method implemented in Banff (Banff Support Team 2008), are designed to handle one constraint at a time. In response pattern (I), the prorating method could proceed as follows: (1) adjust the imputed values for Total costs and Profit with a factor 950/1,030 so that they add up to the observed Turnover, (2) adjust the imputed values for Turnover main and Turnover other with the same factor to satisfy the second edit, and (3) adjust the imputed values of Wages and Other costs, again with the same factor to make them add up to the previously adjusted value of Total costs.

For response pattern (II): step (1) and (2) may be carried out as before, but step (3) needs to be modified unless the observed Wages is to be 'over-written'. Notice that Total costs appears in two edit-rules: $a1$ and $a3$. When the imputed Total costs is only adjusted according to $a1$ in step (1), the relevant information in the observed Wages is ignored. Indeed, depending on the values available it can even happen that Total costs is adjusted downwards in step (1) to the extend that there is no acceptable non-negative solution left for Other costs at step (3). In general, adjusting a variable that appears in multiple edit-rules according to only one of them is not only suboptimal in theory, it also requires an arbitrary choice of the order in which the edit-rules are to be handled, and it may unnecessarily cause a break-down of the procedure.

Under the assumption that the inconsistency is not due to systematic errors, we propose an optimization approach that treats all the constraints simultaneously. To this end it is convenient to express the edit restrictions in matrix notation, as $\mathbf{C}\mathbf{x} = \mathbf{d}$, where $\mathbf{C}$ is the *constraint* (or *restriction*) matrix, and $\mathbf{d}$ a constant vector. For the restrictions $a1 - a3$, we have

$$\mathbf{C} = \begin{pmatrix} 1 & 0 & 0 & 0 & -1 & 0 & 0 & 1 \\ 0 & 0 & -1 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & -1 & 1 \end{pmatrix} \text{ and } \mathbf{d} = \mathbf{0}.$$

The non-zero elements in a *row* of the constraint matrix identify all the variables that are involved in the corresponding edit constraint, and the non-zero elements in a *column* of the constraint matrix identify all the edit constraints that involve the corresponding variable.

In addition, there are often linear inequality constraints. The simplest case is the non-negativity of most economic variables. The constraints can then be formulated as $\mathbf{C}_{eq}\mathbf{x} = \mathbf{d}_{eq}$ and $\mathbf{C}_{ineq}\mathbf{x} < \mathbf{d}_{ineq}$, corresponding to the equality and inequality constraints. For ease of exposition we shall, without noting otherwise, adopt the compact expression $\mathbf{C}\mathbf{x} \leq \mathbf{d}$.

As mentioned earlier, not all the values need or should be adjusted. We therefore make a general distinction between *free* (or adjustable) and *fixed* (not adjustable) variables. This includes as a special case the situation where all the data values are considered adjustable. We emphasize that the distinction is not necessarily that between the imputed and observed variables, and imputation may have been carried out for missing values as well as erroneous observed ones. For instance, some imputed values may be held fixed because they are derived by logical reasoning as in deductive imputation, or they may have been obtained from external sources that are considered more reliable. Whereas some observed values may be considered unreliable and are allowed to be changed. Given the absence of systematic errors, a general approach is to identify the adjustable variables by "error localization" (e.g., de Waal, Pannekoek and Scholtus 2011), treating the imputed and observed values as equally error-prone. Nevertheless, in much of the text below we shall treat the imputed values as adjustable and the observed ones as fixed for ease of elaboration.

Given the free and fixed variables, the complete data record is accordingly partitioned into sub-vectors $\mathbf{x}_{free}$ and $\mathbf{x}_{fixed}$, and the constraints matrix into $\mathbf{C}_{free}$ and $\mathbf{C}_{fixed}$, containing the columns of $\mathbf{C}$ that correspond to $\mathbf{x}_{free}$ and $\mathbf{x}_{fixed}$, respectively. The constraints for the adjustable variables are then given by $\mathbf{C}_{free}\mathbf{x}_{free} \leq \mathbf{d} - \mathbf{C}_{fixed}\mathbf{x}_{fixed}$ or, equivalently,

$$\mathbf{A}\mathbf{x}_{free} \leq \mathbf{b} \tag{1.1}$$

where the matrix $\mathbf{A}$ represents the constraints on the free variables and will be called the *accounting* matrix and $\mathbf{b}$ the constant vector for these constraints. Notice that, while the constraint matrix $\mathbf{C}$ is derived a priori from the edit-rules alone, without reference to the actual data, and is the same for all the records, the accounting matrix $\mathbf{A}$ is generally different from one record to another, since the distinction between free and fixed variables varies across the units.

Our strategy to remedy the micro inconsistency problem in imputed data is to make adjustments to the adjustable values that are minimal according to some chosen distance (or discrepancy) measure, such that the adjusted record satisfies all the edit-rules. All the constraints are simultaneously handled assuming the absence of systematic errors.

The rest of the paper will contain the following. The optimization approach will be outlined in Section 2. We consider different distance (or discrepancy) measures, the adjustments they generate, and illustrate their properties and interpretations using the example record above. In Section 3 we discuss possible extensions of the basic approach to adjustments based on statistical assumptions in addition to logical constraints, treatment of categorical data, unit imputation with adjustments, and adjustments for macro-level benchmarking constraints in combination with micro-level consistency. In Section 4 we

examine the pasture area data from the Norwegian Agriculture Census 2010, including an approach to the assessment of uncertainty due to editing. A final short summary is provided in Section 5.

2 The minimum adjustment approach

2.1 The optimization problem

We propose to resolve the consistency problem outlined above by adjusting the free variables simultaneously and as little as possible, such that all the edit-rules are satisfied. Let the *adjustable* part of the record *before* adjustment be denoted by a J-vector $\mathbf{x}_0$ and by $\tilde{\mathbf{x}}$ the corresponding J-vector *after* the adjustment. The optimization problem can be formulated as:

$$\begin{aligned} \tilde{\mathbf{x}} &= \arg\min_{\mathbf{x}} D(\mathbf{x}, \mathbf{x}_0) \\ \text{s.t.} \quad & \mathbf{A}\tilde{\mathbf{x}} \leq \mathbf{b}, \end{aligned} \tag{2.1}$$

where $D(\mathbf{x}, \mathbf{x}_0)$ is a function measuring the distance (or discrepancy) between $\mathbf{x}$ and $\mathbf{x}_0$, and $\mathbf{A}$ the $K \times J$ accounting matrix associated with the K constraints on $\tilde{\mathbf{x}}$ given in (1.1). We will consider different functions D in Section 2.2.

The conditions for a solution to the minimization problem (2.1) can be found by inspection of the Lagrangian for this problem, which can be written as

$$L(\mathbf{x}, \boldsymbol{\alpha}) = D(\mathbf{x}, \mathbf{x}_0) + \boldsymbol{\alpha}^T (\mathbf{A}\mathbf{x} - \mathbf{b}) \tag{2.2}$$

where $\boldsymbol{\alpha}$ is a K-vector of Lagrange multipliers, or *dual* variables, with components α_k, one for each of the K constraints, and $\mathbf{a}_k$ the k^{th} row (corresponding to constraint k) of the accounting matrix $\mathbf{A}_{K\times J}$. Notice that an additional non-negativity restriction needs to be applied to each α_k corresponding to an inequality constraint, but not the α_k of an equality constraint.

From optimization theory it is well known that for a convex function $D(\mathbf{x}, \mathbf{x}_0)$ and linear constraints, the solution to (2.1) is given by vectors $\tilde{\mathbf{x}}, \tilde{\boldsymbol{\alpha}}$ that satisfy the so-called Karush-Kuhn-Tucker (KKT) conditions (see, e.g., Luenberger 1984; Boyd and Vandenberghe 2004). One of them is that the gradient of the Lagrangian w.r.t. $\mathbf{x}$ is zero when evaluated at $\tilde{\mathbf{x}}, \tilde{\boldsymbol{\alpha}}$, i.e.,

$$L'_{x_j}(\tilde{\mathbf{x}}, \tilde{\boldsymbol{\alpha}}) = D'_{x_j}(\tilde{\mathbf{x}}, \mathbf{x}_0) + \sum_k a_{kj}\tilde{\alpha}_k = 0, \tag{2.3}$$

where a_{kj} is the (k, j)-element of $\mathbf{A}$, and $L'_{x_j}(\tilde{\mathbf{x}}, \tilde{\boldsymbol{\alpha}})$ the gradient of L w.r.t. x_j evaluated at $\tilde{\mathbf{x}}$ and $\tilde{\boldsymbol{\alpha}}$, and D'_{x_j} that of D. From (2.3), we can see how different choices for D lead to different solutions to the adjustment problem, which we will refer to as the *adjustment models*.

2.2 Distance functions and adjustment models

A widely used distance function in many areas of statistics is the weighted least squares (WLS) function given by $D(\mathbf{x}, \mathbf{x}_0) = 1/2\,(\mathbf{x} - \mathbf{x}_0)^T \mathbf{W}(\mathbf{x} - \mathbf{x}_0)$, where $\mathbf{W}$ is a diagonal matrix with diagonal elements w_j, for $j = 1, \ldots, J$. We then obtain, from (2.3), the adjustment model

$$\tilde{x}_j = x_{0,j} - \frac{1}{w_j}\sum_k a_{kj}\tilde{\alpha}_k. \quad (2.4)$$

The WLS-criterion thus results in additive adjustments: the total adjustment to the *initial* value $x_{0,j}$ is the weighted sum of the adjustments that correspond to each of the K constraints. The adjustment due to the k^{th} constraint depends on the following:

- The adjustment parameter (i.e., the dual variable) $\tilde{\alpha}_k$ that describes the amount of adjustment. A smaller value for $\tilde{\alpha}_k$ (in absolute sense if k refers to an equality constraint) corresponds to a smaller adjustment; a zero value for $\tilde{\alpha}_k$ means that no adjustment due to that constraint takes place.
- The constant a_{kj} (i.e., an element of the accounting matrix) describes the direction and size of the adjustment to variable j. Often, a_{kj} is 1, -1 or 0 and then describes whether $x_{0,j}$ is adjusted by $\tilde{\alpha}_k$, $-\tilde{\alpha}_k$ or not at all.
- The weight w_j : variables with larger weights are adjusted less than those with smaller weights. The special case of $w_j \equiv 1$ yields the ordinary least squares (LS) criterion, where the amount of adjustment due to each constraint is the same for all the relevant variables.

A specific choice of the weights is $w_j = 1/x_{0,j}$, for $j = 1,...,J$, in which case the squared relative adjustments are minimized and a larger initial value (i.e., $x_{0,j}$) is adjusted more than a smaller one *in absolute sense*. Dividing (2.4) by $x_{0,j}$ we obtain

$$\frac{\tilde{x}_j}{x_{0,j}} = 1 - \sum_k a_{kj}\tilde{\alpha}_k, \quad (2.5)$$

which is an additive adjustment model for the *ratio* between the adjusted and unadjusted values. It may be noticed that this is the first-order Taylor expansion (i.e., around 0 for all the $\tilde{\alpha}_k$'s) to the multiplicative adjustment given by

$$\frac{\tilde{x}_j}{x_{0,j}} = \prod_k \left(1 - a_{kj}\tilde{\alpha}_k\right). \quad (2.6)$$

From (2.5) we see that $\tilde{\alpha}_k$ determines the relative change from the initial $x_{0,j}$ to the adjusted $\tilde{x}_j$, which in absolute sense is usually much smaller than unity. For instance, $\tilde{\alpha}_k = \pm 0.2$ implies $|20\%|$ adjustment of $x_{0,j}$ if $a_{kj} = \pm 1$, which is large in practice. The products of the $\tilde{\alpha}_k$'s are therefore often much smaller than the α_k's themselves, in which case (2.5) becomes a good approximation to (2.6), and one may regard the WLS adjustment to be roughly given as the product of all the constraint-specific multiplicative adjustments.

Multiplicative adjustment by (2.6) may change the sign of $x_{0,j}$ if $a_{kj}\tilde{\alpha}_k > 1$ for some k. Multiplicative adjustments that preserve the sign of the initial $x_{0,j}$ can be obtained using the

Kullback-Leibler (KL) divergence measure (not formally a distance function), given by $D_{KL} = \sum_j x_j \left(\ln x_j - \ln x_{0,j} - 1\right)$. We then have, from (2.3), the adjustment model

$$\tilde{x}_j = x_{0,j} \prod_k \exp\left(-a_{kj}\tilde{\alpha}_k\right). \tag{2.7}$$

The adjustment due to constraint k is equal to 1 if a_{kj} is 0 (i.e., no adjustment), it is $\exp(\tilde{\alpha}_k)$ if a_{kj} is 1 and it is $1/\exp(\tilde{\alpha}_k)$ if a_{ik} is -1. Since $1 - a_{kj}\tilde{\alpha}_k$ is the first-order approximation of $\exp\left(-a_{kj}\tilde{\alpha}_k\right)$ around $\tilde{\alpha}_k = 0$ if $a_{kj} \pm 1$, the WLS and KL criteria can be expected to yield similar adjustments as long as these are small or moderate.

2.3 Methods for solving the minimum adjustment problem

The general convex optimization problem (2.1) can be solved explicitly if the objective function is the weighted least squares and there are only equality constraints. In this case, the Lagrangian is $L(\mathbf{x}, \boldsymbol{\alpha}) = 1/2\,(\mathbf{x} - \mathbf{x}_0)^T\, \mathbf{W}(\mathbf{x} - \mathbf{x}_0) + \boldsymbol{\alpha}^T(\mathbf{A}\mathbf{x} - \mathbf{b})$, and the equations to be solved are

$$L'_{\mathbf{x}}(\mathbf{x}, \boldsymbol{\alpha}) = \mathbf{W}(\mathbf{x} - \mathbf{x}_0) + \mathbf{A}^T\boldsymbol{\alpha} = \mathbf{0} \tag{2.8}$$

$$L'_{\boldsymbol{\alpha}}(\mathbf{x}, \boldsymbol{\alpha}) = \mathbf{A}\mathbf{x} - \mathbf{b} = \mathbf{0}. \tag{2.9}$$

Solving (2.8) for $\mathbf{x}$ and substituting the result in (2.9) we obtain

$$\tilde{\boldsymbol{\alpha}} = \left(\mathbf{A}\mathbf{W}^{-1}\mathbf{A}^T\right)^{-1}(\mathbf{A}\mathbf{x}_0 - \mathbf{b})$$

and then, on back substitution in (2.8), we obtain explicitly

$$\tilde{\mathbf{x}} = \mathbf{x}_0 - \mathbf{W}^{-1}\mathbf{A}^T\left(\mathbf{A}\mathbf{W}^{-1}\mathbf{A}^T\right)^{-1}(\mathbf{A}\mathbf{x}_0 - \mathbf{b}). \tag{2.10}$$

For other objective functions and with inequality constraints in general, there are no explicit solutions to (2.1). However, there are many free or commercial algorithms for the convex optimization problem. For the application in this paper we used the R programming language and applied the so-called row-action or Successive Projection Algorithms (SPA) - see e.g., Censor and Zenios (1997). The SPA is an iterative algorithm that uses the constraints (rows of the accounting matrix) one by one. In one iteration the x-vector is sequentially adjusted to each of the constraints. The operation of adjusting to a single constraint requires only to update the elements of the x-vector that are involved in that constraint (corresponding to the non-zero elements of the currently processed row of the accounting matrix). After all constraints are visited one iteration is completed and the next one is started. For the WLS criterion, an R-package is available that implements the SPA and is especially designed for the adjustment problem (van der Loo 2012).

2.4 Example revisited

Table 2.1 shows the minimum adjustments of the example record in Table 1.1, using the LS-, WLS- and KL-criterion, respectively. The observed values are treated as fixed and shown in bold, the imputed

values are adjustable. For the WLS method we use $w_j = 1/x_{0,j}$, giving results that are equal to the KL-criterion up to the first decimal.

For both response patterns, the LS adjustment procedure leads to a negative value for Turnover other which is not acceptable (Table 2.1). When the LS-procedure is rerun with a non-negativity constraint for the variable Turnover other, the result is simply a zero for that variable and 950 for Turnover main due to constraint $a2$. Without the non-negativity constraint, the LS-adjustments are -40 for x_3 and x_4, and -16 for x_6 and x_7, i.e., same adjustment for each pair of variables that appear in the same constraint. The variable Total costs (x_8) is part of two constraints and the total adjustment to this variable consists of two additive components. One component is due to constraint $a1$, and the other due to $a3$. For response pattern (I), the first component is -48 and the second component is 16, and the two add up to -32 in Table 2.1.

Table 2.1
Imputation and adjustment of business record in Table 1.1. DI: Partial donor imputation without adjustment; LS: Least-squares distance; WLS: Weighted least-squares distance; KL: Kullback-Leibler divergence measure; GR: Generalized ratio adjustments

Variable	Name	Response (I)				Response (II)			
		DI	LS	WLS/KL	GR	DI	LS	WLS/KL	GR
x_1	Profit	330	282	291	304	330	260	249	239
x_2	Employees	20	20	20	18	**25**	**25**	**25**	**25**
x_3	Turnover Main	1,000	960	922	922	1,000	960	922	921
x_4	Turnover Other	30	-10	28	28	30	-10	28	29
x_5	Turnover	**950**	**950**	**950**	**950**	**950**	**950**	**950**	**950**
x_6	Wages	500	484	470	461	**550**	**550**	**550**	**550**
x_7	Other costs	200	184	188	184	200	140	151	161
x_8	Total costs	700	668	658	646	700	690	701	711

The WLS/KL adjustments are larger, in absolute sense, for larger imputed values than for smaller ones. In particular, the adjustment to Turnover other is only -2.3, so that no negative adjusted value results in this case, whereas the adjustment to Turnover main is -77.7. The multiplicative nature of these adjustments can be observed as the adjustment *factor* for both these variables is 0.92 (for both response patterns). The adjustment factor for Wages and Other costs in response pattern (I) is equally 0.94 because these variables are in the same constraint $a3$, such that the ratio between their initial values is unaffected by this adjustment. However, the initial ratio of each of these variables to Total Costs is not preserved because Total Costs has a different sign in the constraint $a3$ and, moreover, Total Costs is also part of constraint $a1$ so that it is subjected to two adjustment factors.

3 On possible extensions to related adjustment problems

3.1 Generalized ratio adjustments

The ratio model is routinely used for case weighting in business surveys under the assumption that the economic variables can all be related proportionally to a common measure-of-size of the business unit, see

e.g., Särndal, Swensson and Wretman (1992). Motivated by the ratio model one could multiply all the donor values by 950/1,030 to obtain the imputed values for the example record under response pattern (I), including the variable Employees (x_2) for which the initial imputed value 20 does not formally violate any constraints. This shows that there may be situations where, in addition to the logical and accounting constraints, adjustments may be introduced based on statistical assumptions.

For response pattern (II), the observed Employees (x_2), Turnover (x_5) and Wages (x_6) can all potentially be used as the measure-of-size variable in a ratio model, so that a single ratio adjustment does not present itself. However, we may postulate the existence of a *common* ratio between the recipient and donor records under the ratio model, and regard the observed ratios (i.e., 20/25 for Employees, 950/1,030 for Turnover and 550/500 for Wages) as its random manifestations. Then, it seems that a plausible approach is to identify this common ratio as the value that minimizes the variance, or any other dispersion measure that is deemed suitable, of the three individual ratios. Finally, insofar as the common ratio pertains to the other variables, it becomes possible to adjust them using the following *generalized ratio* (GR) approach.

Assume the multiplicative adjustment model $\tilde{x}_j = x_{0,j}\delta_j$, where each δ_j is a random manifestation of a theoretical common ratio. Put the distance function

$$D(\tilde{\mathbf{x}}, \mathbf{x}_0) = 1/2\left(\boldsymbol{\delta}^T\boldsymbol{\delta} - \bar{\delta}^2\right) \tag{3.1}$$

where $\boldsymbol{\delta}$ is the vector of δ_j's and $\bar{\delta}$ the mean of them. For all the variables subjected to the common ratio, including both free and fixed ones, we now carry out the adjustment in two steps. The first step is a conceptual one, where we imagine that an adjustment $\tilde{x}_j / x_{0,j}$ is made to the fixed variables: if $\tilde{x}_j = x_j$ is observed and fixed, then $\delta_j = x_j / x_{0,j}$, whereas $\delta_j = 1$ if $\tilde{x}_j$ is the imputed value $x_{0,j}$ but to be held fixed from 'further' adjustment. At the second step, adjustments are made to the initial values of the free variables by solving the optimization problem (2.1) with (3.1) as the distance function. This yields the GR adjustments of the free variables involved.

An important condition of the GR approach is that at least one of the δ_j's must relate to a fixed variable. Otherwise, $\tilde{x}_j \equiv x_{0,j}$ would be the trivial solution because this always yields $D = 0$. Notice that we have suppressed the denotation J in (3.1), and slightly abused the denotations $\mathbf{x}_0$ and $\tilde{\mathbf{x}}$ introduced for (2.1). Take response pattern (I) in Table 1.1, the fixed value $x_5 = 950$ needs to be included in (3.1), yielding $\delta_5 = \tilde{x}_5 / x_{0,j} = x_5 / x_{0,j} = 950/1,030$. Solving (2.1) for all the other variables yields then $\delta_j \equiv 950/1,030$ and $D = 0$. Whereas, without including δ_5, one would have merely obtained $D = 0$ at $\delta_j = 1$ and $\tilde{x}_j = x_{0,j}$ for $j \neq 5$.

The GR adjustments for response pattern (II) are given in Table 2.1. All the three observed δ_j's, for $j = 2, 5$ and 6, are included in (3.1) and held fixed for the optimization problem. The results are seen to be close to the WLS/KL adjustments. The empirical variance of the multiplicative factors is 0.0270 for the GR adjustments, 0.0276 for WLS/KL and 0.1434 for LS. The relative sum of squared changes, i.e., twice the WLS distance, is 50.6 for the WLS/KL adjustments, 51.6 for GR and 78.0 for LS. Finally, the unweighted sum of squared changes, i.e., twice the LS distance, is 20,925 for the LS adjustments, 23,976 for WLS/KL and 25,090 for GR. Thus, in terms all the three distance functions, the GR adjustments are closer to WLS/KL than LS.

Now, the distance (or discrepancy) measures considered in Section 2.2 may be characterized as *decomposable*, since the overall distance between two vectors is given as a (weighted) sum of the 'distances' between the corresponding components. A consequence is that a variable that does not stand in any constraints will retain the initial value under the minimum adjustment approach. In contrast, the distance (3.1) is *non-decomposable*, where each adjustment is dependent on the other adjustments. As a result even the values that are not explicitly involved in any constraints will be adjusted as long as they are included in the distance function, because of the changes made to the variables that are constraint-bound. The variable Employees provides an example in Table 2.1. The GR approach provides thus a possibility for adjustments based on statistical assumptions in addition to logical and accounting constraints. Indeed, with a single fixed variable included in (3.1), the GR adjustments are reduced to a common proportional adjustment, in accordance with the ratio-adjustment intuition in this case. With multiple fixed variables included, the GR approach aims at a kind of most-uniform adjustments as a generalization of the single-ratio model. For response pattern (II) in Table 1.1, the approach at once takes into account all the three observed ratios. To achieve the same by formulating an explicit statistical model for exactly this response pattern is not as practical in a production setting.

3.2 Adjustments involving categorical data

A categorical variable carries different constraints from a continuous one. It is worth considering the extent to which categorical variables may be incorporated in the optimization approach. We shall distinguish three types of categorical data that are common in practice.

Firstly, we call a categorical/discreet variable *pseudo-continuous* if in practice it can be dealt with as if it were a continuous variable. Typical examples of pseudo-continuous variables are age, number of employees, household size, etc. Pseudo-continuity can affect the choice of adjustment model and distance function. For instance, both additive and proportional adjustments may be acceptable for the number of employees, whereas a proportional adjustment of household size or age seems unnatural. Still, having chosen the adjustment model and distance function, one may handle a pseudo-continuous variable just like a real one. Rounding is necessary afterwards and its effect needs to be monitored.

Secondly, what we call a *nominal* categorical variable indicates whether a unit falls into a particular category. A nominal variable with M categories, labelled $x = 1, 2, \ldots, M$, carry with it the constraint

$$\prod_{m=1}^{M} (\tilde{x} - m) = 0. \tag{3.2}$$

However, the labels (e.g., 1 = tomatoes, 2 = beans, 3 = cucumbers) are not suitable for operations such as addition, multiplication or rounding. Neither is a nominal value 3 more distant to 1 than 2. Therefore, the constraint (3.2) can not be taken into account under the minimum adjustment approach which assumes interval scale measurements. The adjustment of an observed value that does not satisfy (3.2) must be handled by marking it as missing and, then, imputing some admissible as well as suitable value, i.e., just like in case the value is missing to start with.

Thirdly, a variable may be defined to have value zero for the units that are not eligible. Depending on whether the measure is pseudo-continuous or nominal when the unit is eligible, we have a *semi-continuous/-nominal* variable that has a non-zero probability of being zero. The difference to pseudo-continuity above is that a semi-continuous variable may require an additional non-negativity constraint in

the accounting matrix. Consider then a semi-nominal variable. In practical questionnaire design, such a variable is often split in two, say, X_1 and X_2. Let $X_1 = 1$ if the unit is engaged in a certain activity, say, production of greenhouse vegetables, and let $X_1 = 0$ otherwise. Let X_2 be a nominal measure of activity when $X_1 = 1$, and $X_2 = 0$ otherwise. Formally, the logical constraint can be given as

$$(1 - \tilde{x}_1)\,\tilde{x}_2 + \tilde{x}_1 \prod_{m=1}^{M} (\tilde{x}_2 - m) = 0 \tag{3.3}$$

Consider all the possible data patterns, including when a value is missing (indicated by "$-$"):

- $(x_1, x_2) = (-, x_2)$: The value $\tilde{x}_1$ can be deduced provided admissible x_2, i.e., x_2 is either 0 or satisfies (3.2), otherwise the situation turns into case $(x_1, x_2) = (-, -)$ below.
- $(x_1, x_2) = (x_1, -)$: If $x_1 = 0$ then $\tilde{x}_2 = 0$; if $x_1 = 1$ then (3.3) reduces to (3.2) above.
- $(x_1, x_2) = (-, -)$: Both values need to be imputed by values that satisfy (3.3).
- (x_1, x_2) : Violation of (3.3) is e.g., the case if $(x_1, x_2) = (1, 0)$ or if $x_1 = 0$ and $x_2 > 0$. We have case $(-, x_2)$ above if x_2 is fixed, $(x_1, -)$ if x_1 is fixed, or $(-, -)$ if neither is fixed.

To summarize, the constraints (3.2) and (3.3) can not be handled by the minimum adjustment approach with linear constraints considered before. Instead, they need to taken care of by the imputation method. Often, donor-based imputation (e.g., Statistics Canada's CANCEIS software that implements the Nearest Neighbour Imputation Methodology, NIM) can be designed to impute categorical data such that user specified constraints are satisfied, see e.g., Bankier, Lachance and Poirier (2000).

3.3 Adjustment of donor-based unit imputation

In donor-based unit imputation the whole record of values are taken from the chosen donor. This has advantages over joint modelling of all the target variables if there are many of them. Chen and Shao (2000) establish the consistency of survey estimator based on nearest neighbour imputation (NNI) under mild conditions. The key assumption is that the difference in the conditional expectations of any target variable between a donor and a receptor, given the variables on which the distance metric is calculated, is bounded by the "distance" between them. That is, they have the same expectations for all the statistical variables if the "distance" between them is zero.

There is thus a need for adjusting donor-based unit imputation when the "distance" between the receptor and the donor is not zero. To illustrate with the example record in Table 1.1, suppose Turnover (x_5) is always known from the administrative source and is used for donor identification, so that partial imputation under response pattern (I) becomes unit imputation. Since Turnover of the receptor differs from that of the donor, the distance between them is not zero, and it seems natural that the donor values should be adjusted to take this difference into account. Indeed, now that there are constraints involving Turnover, adjustments are necessary in any case.

Let $\mathbf{x}$ contain the variables that may be missing. Let $\mathbf{z}$ contain the known variables that are used for donor identification. Let $\mathbf{x}^* = \left(\mathbf{x}^T, \mathbf{z}^T\right)^T$ be the combined vector of variables. Unit imputation (giving $\mathbf{x}_0$) can be regarded as partial imputation of the missing sub-vector $\mathbf{x}$ of $\mathbf{x}^*$. The need for adjustment of

unit imputation may arise if there are edit-rules that involve both values of $\mathbf{x}$ and $\mathbf{z}$, and/or if the $\mathbf{z}$-values do not match exactly between the donor and receptor. Indeed, unit imputation without adjustment may rather be considered exceptional in practice.

3.4 Macro-level benchmarking in addition to micro-level constraints

A business census requires imputation and editing in order to arrive at a complete dataset for statistical production. Or, a statistical register may be constructed based on a combination of administrative data and one or several sample surveys. Editing and imputation are again necessary. A common feature is that, unlike survey sampling, no case weighting is needed.

When processing such data, macro-level *benchmark* constraints are frequently imposed due to concerns for statistical efficiency and/or macro-level consistency with external sources. A benchmark constraint is satisfied if the complete data add up to the given benchmark total, which may refer to different aggregation levels, i.e., containing both population and sub-population totals. For instance, certain key national totals may be estimated by some suitable method and imposed as benchmark constraints afterwards. Or, a set of domain-level benchmark constraints may be derived by some small area estimation technique. Also benchmark constraints from external sources are common in structural business statistics - an example from the Norwegian Agriculture Census 2010 will be described in Section 4.

Methods for imputation under benchmark constraints have been studied by Beaumont (2005), Chambers and Ren (2004), Zhang (2009) and Pannekoek, Shlomo and de Waal (2013). The approach taken here is similar to the one taken in the first two papers. In both these papers a weighted least squares distance between initial imputed values (or outlying values in the case of Chambers and Ren 2004) and adjusted imputed values is minimized subject to the constraint that sample-weighted totals based on the adjusted data are equal to the benchmark totals. Here, we assume that some suitable imputation method has been applied to yield the initial complete population dataset, which may or may not be benchmarked. The inconsistency problem on the micro-level implies that adjustments of the initial complete data set will be necessary in general.

Denote by $\mathbf{X}$ the complete dataset of interest, where each row corresponds to a unit-level record as the one in Table 1.1, and each column corresponds to a particular variable. Let $\mathbf{X}_0$ be the initial complete dataset after imputation and $\tilde{\mathbf{X}}$ the adjusted dataset. Each benchmark constraint applies to a particular column vector of $\mathbf{X}$ and over the units that fall under its domain. That is, it can be expressed generically as $\mathbf{r}^T \mathrm{col}(\mathbf{X}) = t$, where $\mathrm{col}(\mathbf{X})$ is the column vector of concern, and $\mathbf{r}$ is the indicator vector for whether a unit belongs to the domain of concern, and t the benchmark total. In this way all the benchmark constraints may be summarized as

$$[\mathbf{r}]^T [\mathrm{col}(\mathbf{X})] = \mathbf{t} \tag{3.4}$$

where each column of $[\mathrm{col}(\mathbf{X})]$ corresponds to a benchmark constraint, and each column of $[\mathbf{r}]$ the corresponding indicator vector, and $\mathbf{t}$ the vector of all the benchmark totals. Notice the similarity between (3.4) and (1.1). A minimum adjustment approach follows on specifying the adjustable and fixed values and the distance (or discrepancy) function.

Both the benchmark constraints and the micro-level constraints can be seen as linear constraints on the very long vector containing all elements of $\mathbf{X}$, $\text{vec}(\mathbf{X})$, say. Conceptually, all constraints together can therefore be expressed in the form (1.1). The restriction matrix of this formulation is, however, huge and very sparse. The rows corresponding to the micro-level constraints contain possibly non-zero values corresponding to the values in the record they apply to and zeros for all other values of $\text{vec}(\mathbf{X})$ and the rows corresponding to the benchmark constraints contain non-zero elements only corresponding to the values in $\text{vec}(\mathbf{X})$ that contribute to that benchmark total. In practice, the optimization problem generated by (3.4) in addition to the micro-level constraints can be handled using the SPA, i.e., one constraint at a time and operating only on the elements of $\text{vec}(\mathbf{X})$ corresponding to the non-zero elements in that constraint, without actually forming this huge and sparse constraint matrix. For the benchmark constraints we only need to process the columns of $[\text{col}(\mathbf{X})]$ one by one and for the micro-level constraints we process each unit-level record one at a time. These iterative minimum adjustments along the columns and rows of $\mathbf{X}$ resemble the iterative proportional fitting (or raking) algorithm for fitting log-linear models to contingency table data and for adjusting (contingency) tables to new margins, which is formally identical to a SPA with the KL-divergence and equality constraints only.

4 Case study

4.1 Imputation and adjustment of pasture data

The population for the "main questionnaire" of the Norwegian Agriculture Census 2010 contains about 45,000 units. Questions 22 - 24 deal with pasture area:

- Question 22 inquires the units that possess productive pasture.
- Question 23 inquires the total productive pasture area in 2010.
- Question 24 inquires the composition of pasture area by the last time it was seeded: (1) 2006 - 2010, (2) 2001 - 2005, and (3) 2000 or earlier.

Denote by $x_{0,1}, x_{0,2}$ and $x_{0,3}$ the three reported categories of pasture area in Question 24. Let $x_0 = \sum_{j=1}^{3} x_{0,j}$ be the sum that is the subject of Question 23. Now, this total is available from the government agency that administers the relevant subsidy. In editing the reported x_0 is overwritten by the administrative figure, denoted by $\tilde{x}$, and held as fixed afterwards. Next, Question 22 can be inferred given $\tilde{x}$ and held as fixed afterwards, so that only Question 24 remains to be handled.

Below we describe the treatment of the 34,480 units that have productive pasture area according to their respective observation patterns (Table 4.1, where the unit index i of all the variables was omitted for ease of presentation).

- 10,378 units reported a total pasture area that is consistent with the administrative source: these are the potential donors; no adjustment is needed.
- 11,827 units have a reported total that is greater than the known value: these have a micro-level inconsistency problem. Of course, missing values can also be the case if $\sum_j r_j < 3$, but the

chance is small, so we shall assume that there are no missing values among these units. All the observed values are adjustable, such that the accounting equation is given by

$$\sum_{j;r_j=1} \tilde{x}_j = \tilde{x}.$$

The GR approach simply yields the proportional adjustment $\tilde{x} \Big/ \sum_{j;\, r_j=1} x_{0,j}$. The same adjustment is given by the WLS-approach with $w_j = 1 \Big/ x_{0,j}$ if $r_j = 1$, as well as by the KL approach. We notice that there is no particular motivation for considering additive adjustments for these data.

- 3,876 units have *no* reported pasture area of any kind, despite they have productive pasture area according to the administrative source: these constitute unit-missing records. The nearest-neighbour (NN) donor is found according to $\tilde{x}$, within each of the 12 "farming forms", which is a classification known for the whole population. In the case of multiple NN donors, we choose the one with the shortest physical distance, which make the NN-imputation completely deterministic, given all the $\tilde{x}$-values. Finally, a proportional adjustment of the donor values is carried out in order to satisfy the accounting equation

$$\sum_{j;r_j^*=1} \tilde{x}_j = \tilde{x}$$

where r_j^* is the observation/reporting indicator associated with the donor.

- 3,019 units have reported pasture areas of *all* the three kinds, but their sum is less than the known total: these have a micro-level inconsistency problem. A proportional adjustment is applied to all the reported values w.r.t. the accounting equation $\sum_{j=1}^{3} \tilde{x}_j = \tilde{x}$.

- The last two groups are the 2,703 units with one kind of reported pasture area and the 2,677 units with two kinds of reported pasture area. Obviously, that the reported total is less than the known value here may be caused by inconsistency and/or missing values. To avoid introducing systematic pattern through editing, we let the decision depend on the donor. Take a unit with only one reported pasture area. Firstly, the potential donors are limited to those from the same "farming form", as well as having *at least* the same kind of pasture area. The NN donor is then selected among these to minimize

$$\max\left(\left|\tilde{x}^* \Big/ \tilde{x} - 1\right|, \left|x_j^* \Big/ \tilde{x}^* - x_{0,j} \Big/ \tilde{x}\right|_{j;r_j=1}\right)$$

where $\left(x_1^*, x_2^*, x_3^*\right)$ and $\tilde{x}^*$ are the values of the potential donor. In other words, the NN donor is selected both w.r.t. the relative difference between the total pasture area as well as the proportion of the reported kind of pasture area to the corresponding total. Let the NN donor be associated with $\mathbf{x}^*$ and $\mathbf{r}^*$. If $\sum_j r_j^* > 1 = \sum_j r_j$, then we assume that there are missing values where $r_j^* = 1$ but $r_j = 0$; whereas, if $\sum_j r_j^* = \sum_j r_j$, then we assume that there is only an inconsistency problem. The remaining imputation and adjustment actions are straightforward. The same treatment is applied to the units with two reported pasture areas, with obvious modifications due to $\sum_j r_j = 2$.

Table 4.1
Observation pattern among units with productive pasture area: $r_j = 1$ if $x_{0,j}$ is reported, $r_j = 0$ otherwise; $j = 1, 2, 3$ for three categories of pasture area

Total	$\sum_j r_j x_{0,j} = \tilde{x}$	$\sum_j r_j x_{0,j} > \tilde{x}$	$\sum_j r_j x_{0,j} < \tilde{x}$			
			$\sum_j r_j = 0$	$\sum_j r_j = 1$	$\sum_j r_j = 2$	$\sum_j r_j = 3$
34,480	10,378	11,827	3,876	2,703	2,677	3,019

The sub-population and population totals based on imputation with adjustments are given in Table 4.2, in comparison with raw data totals and the census file totals. We notice the following. (a) The census file had been edited in a 'traditional' way that involves much clerical work (about 1.5 man-year in total). In contrast, the editing procedures here are fully automated, and everything (i.e., exploratory analysis, decision of the treatments, programming and processing) was done in less than two days. Although the questions concerning pasture areas are only 3 out of a total of 36 questions of the "main questionnaire", it is obvious that the potential saving in time could be enormous. (b) The differences between the imputed totals and the census totals are small for all sub-populations, compared to those between the raw data and the census totals. All the changes from the raw data are in the 'right' direction, judged by the census results. One may conclude that the automated editing procedures have achieved most of the census editing results. (c) It is possible to introduce benchmark constraints in addition. An an illustration, we used the census file sub-population totals for the 3,876 unit-missing records, in addition to the known pasture area total for each of them. Convergence was reached in 23 iterations with the WLS criterion. (d) For the 5,380 units where partial missing may be the case, imputation of 'missing' values was carried for about 25% of them in the census processing, whereas it is about 75% by the editing procedure here. The number of cases for partial missing is probably under-estimated in the census file because it is based on selective manual checks. In any case, not withstanding the differences in the individual treatments, the edited totals are fairly close to each (Table 4.2, under $0 < \sum_j r_j < 3$).

Table 4.2
Sub-population and population pasture area totals based on raw data, imputation with adjustments and census production data. (All figures $\times 10^5$)

	$\sum_j r_j x_{0,j} > \tilde{x}$			$\sum_j r_j x_{0,j} < \tilde{x}$					
				$\sum_j r_j = 3$			$0 < \sum_j r_j < 3$		
Raw	8.20	6.95	12.76	1.40	1.45	1.53	1.33	0.86	3.05
Impute & Adjust	5.24	4.34	8.71	1.72	1.81	1.88	2.01	1.87	3.51
Census	5.47	4.37	8.45	1.73	1.85	1.84	2.04	1.54	3.80
	$\sum_j r_j = 0$			$\sum_j r_j > 0$			**Total**		
Raw	-	-	-	14.0	12.4	21.9	-	-	-
Impute & Adjust	1.20	1.06	1.93	12.2	11.3	19.3	13.43	12.38	21.17
Census	1.31	1.23	1.66	12.6	11.0	19.1	13.95	12.25	20.79

4.2 Approximate mean squared error estimation

As the measure of uncertainty for the pasture area data here, we use the mean squared error of prediction (MSEP) given by

$$\text{MSEP}_j = E\left\{\left(\tilde{X}_j - X_j\right)^2 \middle| \mathbf{R}_U, \tilde{\mathbf{X}}_U\right\}$$

where $X_j = \sum_{i \in U} x_{ij}$ is the target population total and $\tilde{X}_j = \sum_{i \in U} \tilde{x}_{ij}$ is the corresponding total based on imputation with adjustments, for $j = 1, 2, 3$. Moreover, $\tilde{\mathbf{X}}_U = (\tilde{x}_i)_{i \in U}$ contains the known pasture area totals in the population, and $\mathbf{R}_U$ is the matrix of missing indicators whose i^{th} row is given by (r_{i1}, r_{i2}, r_{i3}).

Now, while it is customary that adjustments due to inconsistency in the micro data are referred to as imputation in statistical data editing, the eventual uncertainty associated with this is generally 'ignored' afterwards. This amounts to assume that $\tilde{x}_{ij} = x_{ij}$ if $r_{ij} = 1$. What remains to be accounted for is the uncertainty associated with the imputation of the missing values and the subsequent adjustment of the donor values, under the assumption that neither imputation nor adjustment introduces bias to the final value. This amounts to assume that $E(\tilde{x}_{ij} - x_{ij}) = 0$ if $r_{ij} = 0$. Under these two assumptions, we have

$$\begin{aligned} \text{MSEP}_j &= E\left\{\left(\sum_{i \in U}(1 - r_{ij})\,\tilde{x}_{ij} - \sum_{i \in U}(1 - r_{ij})\,x_{ij}\right)^2\right\} \\ &= V\left(\sum_{i \in U;\mathbf{r}_i = \mathbf{1}, d_{ij} \geq 1} d_{ij}\delta_{ij}x_{ij}\right) + V\left(\sum_{i \in U; r_{ij} = 0} x_{ij}\right) \\ &\approx \sum_{i \in U;\mathbf{r}_i = \mathbf{1}, d_{ij} \geq 1} d_{ij}^2 V(\delta_{ij}x_{ij}) + \sum_{i \in U; r_{ij} = 0} V(x_{ij}) \end{aligned}$$

where d_{ij} is the number of times x_{ij} is used as a donor value for imputation of missing data, and the decomposition of variance holds provided the distributions of the units are independent of each other. Moreover, provided $d_{ij} \geq 1$,

$$\delta_{ij} = \sum_{k \in U; x^*_{kj} = x_{ij}} \tilde{x}_{kj} \Big/ (d_{ij}x_{ij})$$

where $x^*_{kj} = x_{ij}$ means that x_{ij} is used as the donor value for x_{kj}, and $\tilde{x}_{kj}$ is the final value after adjustment. In other words, δ_{ij} is the combined adjustment made to $d_{ij}x_{ij}$, where $d_{ij}x_{ij}$ would have been the contribution of x_{ij} to $\tilde{X}_j$ through imputation if it had been donor imputation *without* adjustment. Notice that d_{ij} can be treated as a constant in the last (approximate) equation as long as the donor identification depends only on $\mathbf{R}_U$ and $\tilde{\mathbf{X}}_U$. This is true for the 3,876 unit-missing records, but not exactly for the 5,380 units that may have partial missing. As explained in Section 4.1, the NN-identification in fact also depends on the observed x_{ij} - values. For this reason, the last equation holds only approximately.

A ratio model for the conditional variance of x_{ij} seems natural here, i.e.,

$$x_{ij} = \beta_j x_i + \varepsilon_{ij} \text{ where } E(\varepsilon_{ij}) = 0 \text{ and } V(\varepsilon_{ij}) = \sigma_j^2 x_i^{\alpha_j}$$

where $(\beta_j, \sigma_j^2, \alpha_j)$ may vary according to the *composition* of the pasture areas, denoted by $\mathbf{q} = (1,1,1), (1,1,0), (1,0,1)$ and $(0,1,1)$, where $q_{ij} = 1$ if unit i has the j^{th} type pasture and 0

otherwise. Notice that, in the case of $\sum_j q_{ij} = 1$, we have $x_{ij} = \tilde{x}$ if $q_{ij} = 1$, so that the conditional variance is zero. The parameters of this ratio model can be estimated from the 10,378 potential donors satisfying $\sum_j r_j x_{0,j} = \tilde{x}$. Exploratory data analysis shows that $\alpha_j = 2$ is a reasonable choice in all the cases, so that in the calculations below only β_j and σ_j^2 vary according to the observation pattern, denote by $\left(\beta_{j;h}, \sigma_{j;h}^2\right)$ for $h = 1, \ldots, 4$. Notice that, as a result of $\alpha_j \equiv 2$, the same $\hat{\sigma}_{j;h}^2$ will be obtained regardless of j whenever $\sum_j q_{ij} = 2$. Take e.g., $\mathbf{q} = (1,1,0)^T$, we have $\hat{\beta}_1 + \hat{\beta}_2 = 1$, such that the 'standardized' fitted residuals are given by $\hat{\varepsilon}_{i1} / \tilde{x}_i = x_{i1} / \tilde{x}_i - \hat{\beta}_1$ and $\hat{\varepsilon}_{i2} / \tilde{x}_i = x_{i2} / \tilde{x}_i - \hat{\beta}_2 = \left(\tilde{x}_i - x_{i1}\right) / \tilde{x}_i - \left(1 - \hat{\beta}_1\right) = -\hat{\varepsilon}_{i1} / \tilde{x}_i$. In any case, we obtain $\hat{V}_h\left(x_{ij}\right) = \hat{\sigma}_{j;h}^2 \tilde{x}_i^2$ for unit i with composition h.

The adjustment factor δ_{ij} seems difficult to model in advance. But its mean and variance can be estimated empirically *after* imputation and adjustment have been carried out, denoted by $\mu_\delta = E\left(\delta_{ij}\right)$ and $\sigma_\delta^2 = V\left(\delta_{ij}\right)$, respectively. Moreover, we assume δ_{ij} to be independent of x_{ij} conditional on $\tilde{x}_i$. This seems a plausible assumption, since the former depends mostly on how x is distributed in the 'neighbourhood' of $x = \tilde{x}$, whereas the latter depends on the variation across j given that the sum is equal to $\tilde{x}$. For instance, asymptotically as the chance of finding a donor in any arbitrarily close neighbourhood tends to unity, the adjustment factor δ_{ij} tends to 1 in probability, irrespective of the values of x_{ij}. It now follows that, given composition h, an estimate of the corresponding $V_h\left(\delta_{ij} x_{ij}\right)$ is given by

$$\hat{V}_h\left(\delta_{ij} x_{ij}\right) = \hat{\sigma}_{j;h}^2 \tilde{x}_i^2 \hat{\sigma}_\delta^2 + \left(\hat{\beta}_{j;h} \tilde{x}_i\right)^2 \hat{\sigma}_\delta^2 + \hat{\sigma}_{j;h}^2 \tilde{x}_i^2 \hat{\mu}_\delta^2.$$

Finally, combining all the above, we obtain an approximate MSEP estimate as

$$\widehat{\mathrm{MSEP}}_j \approx \sum_h \sum_{i \in U_h; \mathbf{r}_i = \mathbf{1}} d_{ij}^2 \hat{V}_h\left(\delta_{ij} x_{ij}\right) + \sum_h \sum_{i \in U_h; r_{ij} = 0} \hat{V}_h\left(x_{ij}\right).$$

The results of approximate variance estimation are given in Table 4.3. We know in advance that the regression coefficient of the ratio model must vary according to the composition of pasture area, but the estimates of $\sigma_{j;h}^2$ suggest that it has been sensible to allow the variance parameter to depend on h. The estimated mean of δ_{ij} is close to unity for all the pasture area types, making no indications that the assumptions regarding the adjustment factors are unreasonable. The variance of δ_{ij} is clearly the largest for $j = 2$, which is also reflected in the fact that the estimated MSEP here has the largest increase compared to NN-imputation without adjustment. The relative root MSEPs are too small to account for the actual differences between the census totals and the imputed totals (given in Table 4.2). This serves to illustrate the following general impression regarding the assessment of uncertainty due to editing. Systematic effects in terms of the first-order moments of the resulting statistics usually dominate the overall uncertainty due to editing. But they are also more difficult to quantify compared to the second-order variance properties. In the case here, this concerns the two 'first-order' assumptions made in the beginning, i.e., $\tilde{x}_{ij} = x_{ij}$ if $r_{ij} = 1$ and $E\left(\tilde{x}_{ij} - x_{ij}\right) = 0$ if $r_{ij} = 0$. More sophisticated assumptions about the error-mechanism of consistency adjustments in editing are needed in order to progress beyond such an 'optimistic' approach.

Table 4.3
Approximate variance estimation for imputation with adjustment. RMSEP: Root MSEP. RMSEP by NN-imputation without adjustment in parentheses

		$j = 1$	$j = 2$	$j = 3$
$\hat{\beta}_j$	$\mathbf{q} = (1,1,1)$	0.312	0.359	0.329
	$\mathbf{q} = (1,1,0)$	0.346	0.654	-
	$\mathbf{q} = (1,0,1)$	0.407	-	0.593
	$\mathbf{q} = (0,1,1)$	-	0.567	0.433
$\hat{\sigma}_j^2$	$\mathbf{q} = (1,1,1)$	0.0248	0.0511	0.0364
	$\mathbf{q} = (1,1,0)$	0.0478	0.0478	-
	$\mathbf{q} = (1,0,1)$	0.0464	-	0.0464
	$\mathbf{q} = (0,1,1)$	-	0.0798	0.0798
	$(\hat{\mu}_\delta, \hat{\sigma}_\delta^2)$	(0.992, 0.0248)	(1.020, 0.0994)	(1.003, 0.0236)
	$\widehat{\text{RMSEP}}$	3,267 (3,134)	4,190 (3,530)	3,111 (2,925)
	$\widehat{\text{RMSEP}} / \sum_{i;r_{ij}=0} \tilde{x}_{ij}$	1.41%	1.79%	0.93%
	$\widehat{\text{RMSEP}} / \tilde{X}_j$	0.24%	0.34%	0.15%

5 Summary

In this paper we have formulated an optimization approach to the micro-level inconsistency problem that may be caused by measurement errors and/or imputation of missing values. This provides a general methodology that extends beyond the traditional single-constraint adjustment methods such as prorating. All constraints are handled simultaneously; if a variable appears in more than one constraint then it is adjusted according to all of them. Besides being optimal according to the chosen distance (or discrepancy) function, the approach also has the practical advantage that there is no need to specify the order in which the constraints are to be applied.

Several distance (or discrepancy) functions are analysed. It is shown that minimizing the weighted least squares leads to additive adjustments and minimizing the Kullback-Leibler divergence measure leads to multiplicative adjustments. However, for a specific choice of weights the WLS solution of the optimization problem is an approximation to the KL solution.

Adjustments based on statistical assumptions in addition to the logical constraints is introduced under the generalized ratio approach. The GR adjustments can be considered as a generalization of the single-ratio adjustment under a ratio model. All the observed variable-specific ratios between the receptor and donor records are utilized; a variable that does not stand in any constraint can also be adjusted if it is included in the distance function.

Also discussed are adjustments involving categorical data, unit-missing records and macro-level benchmark constraints in addition to the micro-level consistency constraints. Taken together, the proposed optimization approach is applicable to continuous data in a number of situations.

Acknowledgements

The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands.

References

Banff Support Team (2008). *Functional Description of the Banff System for Edit and Imputation*. Technical Report, Statistics Canada.

Bankier, M., Lachance, M. and Poirier, P. (2000). *2001 Canadian Census Minimum Change Donor Imputation Methodology*. Working paper 17, UN/ECE Work Session on Statistical Data Editing, Cardiff.

Beaumont, J.-F. (2005). Calibrated imputation in surveys under a quasi-model-assisted approach. *Journal of the Royal Statistical Society, Series B* (Statistical Methodology), 67, 445-458.

Boyd, S., and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, Cambridge.

Censor, Y., and Zenios, S.A. (1997). *Parallel Optimization*. Theory, Algorithms, and Applications. Oxford University Press, New York.

Chambers, R.L., and Ren, R. (2004). Outlier robust imputation of survey data. In *Proceedings of the Survey Research Methods Section,* American Statistical Association, 3336-3344.

Chen, J., and Shao, J. (2000). Biases and variances of survey estimators based on nearest neighbour imputation. *Journal of Official Statistics*, 16, 113-132.

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

Luenberger, D.G. (1984). *Linear and Nonlinear Programming, Second Edition*. Addison-Wesley, Reading.

Pannekoek, J., Shlomo, N. and de Waal, T. (2013). Calibrated imputation of numerical data under linear edit restrictions. *Annals of Applied Statistics*, 7, 1983-2006.

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

van der Loo, M. (2012). rspa: Adapt numerical records to (in)equality restrictions with the Successive Projection Algorithm. R package version 0.1-5. Available at: http://cran.r-project.org/web/packages/rspa/index.html.

Zhang, L.-C. (2009). *A Triple-Goal Imputation Method for Statistical Registers*. Working paper 28, UN/ECE Work Session on Statistical Data Editing, Neuchâtel, Switzerland.

Survey Methodology, June 2015
Vol. 41, No. 1, pp. 145-164
Statistics Canada, Catalogue No. 12-001-X

Dealing with non-ignorable nonresponse in survey sampling: A latent modeling approach

Alina Matei and M. Giovanna Ranalli[1]

Abstract

Nonresponse is present in almost all surveys and can severely bias estimates. It is usually distinguished between unit and item nonresponse. By noting that for a particular survey variable, we just have observed and unobserved values, in this work we exploit the connection between unit and item nonresponse. In particular, we assume that the factors that drive unit response are the same as those that drive item response on selected variables of interest. Response probabilities are then estimated using a latent covariate that measures the *will to respond to the survey* and that can explain a part of the unknown behavior of a unit to participate in the survey. This latent covariate is estimated using latent trait models. This approach is particularly relevant for sensitive items and, therefore, can handle non-ignorable nonresponse. Auxiliary information known for both respondents and nonrespondents can be included either in the latent variable model or in the response probability estimation process. The approach can also be used when auxiliary information is not available, and we focus here on this case. We propose an estimator using a reweighting system based on the previous latent covariate when no other observed auxiliary information is available. Results on its performance are encouraging from simulation studies on both real and simulated data.

Key Words: Unit nonresponse; Item nonresponse; Latent trait models; Response propensity; Rasch models.

1 Introduction

Nonresponse is an increasingly common problem in surveys. It is a problem because it causes missing data and, more importantly, because such gaps are a potential source of bias for survey estimates. In the presence of unit nonresponse, it is often assumed that each unit in the population has an associated probability to respond to the survey. Such a response probability is unknown and several methods are proposed to estimate it either explicitly, using response propensity modeling like logistic regression models (see e.g., Kim and Kim 2007), or implicitly, using response homogeneity groups or more generally calibration (see Särndal and Lundström 2005, for an overview). Once estimates are computed, a commonly used method to deal with unit nonresponse is reweighting: sampling weights of the respondents are adjusted by the inverse of the estimated response probability providing new weights. Estimation of response probabilities typically requires the availability of auxiliary information, either in the form of the value of some auxiliary variables for all units in the originally selected sample or of their population mean or total.

In this paper, we are particularly interested in the case where the missing data mechanism is non-ignorable, because nonresponse depends on characteristics of interest that are either observed only on the respondents or are completely unobserved, which leads to data that are Not Missing At Random (NMAR). This is typical of, but not limited to, surveys with sensitive questions (concerning drug abuse, sexual attitudes, politics, income, etc). Various approaches are proposed in the survey sampling literature to deal with non-ignorable nonresponse. These approaches can be roughly divided into likelihood based methods and reweighting methods. Note that all of these methods make use of observed auxiliary information. Survey problems with non-ignorable nonrespondents are discussed e.g., in Greenlees, Reece and

1. Alina Matei, Institute of Statistics, University of Neuchâtel, Pierre à Mazel 7, 2000, Neuchâtel, Switzerland. E-mail: alina.matei@unine.ch and Institute of Pedagogical Research and Documentation Neuchâtel, Switzerland; M. Giovanna Ranalli, Dept. of Political Sciences, University of Perugia, Italy. E-mail: giovanna.ranalli@stat.unipg.it.

Zieschang (1982), Little and Rubin (1987), Beaumont (2000), Qin, Leung and Shao (2002), Zhang (2002). Copas and Farewell (1998) introduce into the British National Survey of Sexual Attitudes and Lifestyles a variable called 'enthusiasm-to-respond' to the survey, which is expected to be related to probabilities of unit and item response. A method is proposed that estimates these probabilities using this variable to achieve unbiased estimates of population parameters. An approach based on the use of latent variables for modeling nonignorable nonresponse is given in Biemer and Link (2007), extending the ideas in Drew and Fuller (1980) and using a discrete latent variable based on call history data available for all sample units. The latent variable is computed using some indicators of level of effort based on call attempts.

We propose here a method of reweighting to reduce nonresponse bias in the case of non-ignorable nonresponse. The method does not require the availability of auxiliary information, on the sample or population level, but different assumptions are made. First, it is assumed that item nonresponse is present in the survey and that it affects m variables of particular interest. Thus a response indicator can be defined for each variable ℓ, for $\ell = 1, \ldots, m$, taking value 1 if item ℓ is observed on unit k and 0 otherwise. Next, the response indicators are assumed to be manifestations of an underlying continuous scale which determines a latent variable that is related to the response propensity of the units and to the variable of interest. It is possible to compute such a latent variable for all units in the sample, not only for the respondents, and thus to use it as an auxiliary variable in a response probability estimation procedure. The outcome of this estimation procedure can finally be used in a reweighting fashion.

The use of continuous latent variables to model item nonresponse is considered in Moustaki and Knott (2000). In this paper, we take a different perspective and use latent variable models to address non-ignorable unit nonresponse. We propose to use a latent variable called here 'will to respond to the survey', which is expected to be related to the probability of unit response, similar to the case of the 'enthusiasm-to-respond' variable as defined by Copas and Farewell (1998). Following Moustaki and Knott (2000), '*weighting through latent variable modeling is expected to perform well under non-ignorable nonresponse where conditioning on observed covariates only is not enough*.' Moreover, in the absence of any covariate, we expect that an estimator based on the proposed weighting system using latent variables will perform better in terms of bias reduction than the naive estimator computed on the set of respondents. Moustaki and Knott (2000) propose a reweighting system for *item non-response* using covariates and one or more latent variables. Our major contribution over the existing literature is to construct a weighting system to deal with *unit and item non-response* based only on latent variables and that can also be used in the absence of any other covariate. On the other hand, our approach is different to that of Copas and Farewell (1998), because they survey their 'enthusiasm-to-respond' variable on the respondents to quantify the interest in answering the survey and a set of covariates, while we infer it from the data.

The paper is organized as follows. Section 2 introduces the survey framework and notation. Section 3 illustrates estimation of response probabilities. Section 4 describes the latent trait model used to this end. The proposed estimator and its variance estimation are shown in Section 5. In Section 6, the empirical properties of the proposed estimator are evaluated via simulation studies. In Section 7 we summarize our conclusions.

2 Framework

Let U be a finite population of size N, indexed by k from 1 to N. Let s denote the set of sample labels, so that $s \subset U$, drawn from the population using a probabilistic sampling design $p(s)$. The

sample size is denoted by n. Let $\pi_k = \sum_{s;s\ni k} p(s)$ be the probability of including unit k in the sample. It is assumed that $\pi_k > 0, k = 1,\ldots,N$. Not all units selected in s respond to the survey. Denote by $r \subseteq s$ the set of respondents, and by $\bar{r} = s \setminus r$ the set of nonrespondents. The response mechanism is given by the distribution $q(r|s)$ such that for every fixed s we have

$$q(r|s) \geq 0, \text{ for all } r \in \mathcal{R}_s \text{ and } \sum_{s\in\mathcal{R}_s} q(r|s) = 1, \text{ where } \mathcal{R}_s = \{r|r \subseteq s\}.$$

Under unit nonresponse we define the response indicator $R_k = 1$ if unit $k \in r$ and 0 if $k \in \bar{r}$. Thus $r = \{k \in s|R_k = 1\}$. We assume that these random variables are independent of one another and of the sample selection mechanism (Oh and Scheuren 1983). Since only the units in r are observed, a response model is used to estimate the probability of responding to the survey of a unit $k \in U$, $p_k = P(k \in r|k \in s) = P(R_k = 1|k \in s)$, which is a function of the sample and must be positive.

Suppose that in the survey there are m variables of particular interest. Each respondent is exposed to these m questionnaire variables, labelled $\ell = 1,\ldots,m$. Suppose that the goal is to estimate the population total of some variables of interest and, in particular, of the variable of interest y_j, i.e., $Y_j = \sum_{k=1}^{N} y_{kj}$, with y_{kj} being the value taken by y_j on unit k. In the ideal case, if the response distribution $q(r|s)$ is known, then the p_k's would be known and available to estimate Y_j using a reweigthing approach. Suppose also that item nonresponse is present for variable y_j. Let $r_j = \{k \text{ answers } y_j|k \in r\}$ be the set of respondents for variable y_j. As in the case of unit nonresponse we assume that the units in r_j respond independently of each other. Let $q_{kj} = P(k \text{ answers } y_j|k \in r)$. The final set of weights to be used into a fully reweighting approach to handle unit and item nonresponse is given by $1/(\pi_k p_k q_{kj})$, for all $k \in r_j$, assuming $q_{kj} > 0$. These weights can be for example used in a three-phase fashion in the following Horvitz-Thompson (HT) estimator

$$\hat{Y}_{j,pq,\text{true}} = \sum_{k\in r_j} \frac{y_{kj}}{\pi_k p_k q_{kj}}, \tag{2.1}$$

(see Legg and Fuller 2009, for the properties of estimators under three-phase sampling).

Usually, p_k and q_{kj} are unknown and should be estimated. A nonresponse adjusted estimator is then constructed by replacing p_k and q_{kj} with estimates $\hat{p}_k$ and $\hat{q}_{kj}$ in (2.1). The following sections provide details with this regard.

3 Estimating response probabilities

3.1 Using logistic regression to estimate p_k

Different methods to estimate p_k are proposed in the literature. All of these methods are based on the use of auxiliary information known on the population or sample level. In the case of non-ignorable nonresponse, the variable of interest is itself the cause (or one of the causes) of the response behavior, and a covariance between the former and the response probability is produced through a direct causal relation

(see Groves 2006). In such a case, the response probability p_k could be modeled for $k \in s$ using logistic regression as follows

$$p_k = P\left(R_k = 1 \middle| y_{kj}\right) = \frac{1}{1 + \exp\left(-\left(a_0 + a_1 y_{kj}\right)\right)}, \tag{3.1}$$

or as follows

$$p_k = P\left(R_k = 1 \middle| y_{kj}, \mathbf{z}_k\right) = \frac{1}{1 + \exp\left(-\left(a_0 + a_1 y_{kj} + \mathbf{z}_k'\boldsymbol{\alpha}\right)\right)}, \tag{3.2}$$

where $\mathbf{z}_k = (z_{k1}, \ldots, z_{kt})'$ is a vector with the values taken by $t \geq 1$ covariates on unit k, and a_0, a_1, and $\boldsymbol{\alpha}$ are parameters.

Nonresponse bias in the unadjusted respondent total of the variable of interest y_j depends on the covariance between the values y_{kj} and p_k (see Bethlehem 1988). An example of a covariate that reduces the covariance between y_{kj} and p_k is the interest in the survey topic, such as knowledge, attitudes, and behaviors related to the survey topic (see Groves, Couper, Presser, Singer, Tourangeau, Acosta and Nelson 2006). The set of covariates $\mathbf{z}_k$ could be also related to the variable of interest y_j to reduce sampling variance (Little and Vartivarian 2005).

Since y_{kj} is only observed on respondents, Models (3.1) and (3.2) cannot be estimated. Therefore, usually, the values of $\mathbf{z}_k$ that are known for both respondents and nonrespondents and are related to the y_{kj}'s by a 'hopefully strong regression' (Cassel, Särndal and Wretman 1983) are used in the following model

$$p_k = P\left(R_k = 1 \middle| \mathbf{z}_k\right) = \frac{1}{1 + \exp\left(-\left(a_0 + \mathbf{z}_k'\boldsymbol{\alpha}\right)\right)}. \tag{3.3}$$

Then, maximum likelihood can be used to fit Model (3.3) using the data $(R_k, \mathbf{z}_k)$ for $k \in s$. This leads to estimate $\hat{a}_0$ and $\hat{\boldsymbol{\alpha}}$ and to the estimated response probabilities $\hat{p}_k = 1/\left[1 + \exp\left(-\left(\hat{a}_0 + \mathbf{z}_k'\hat{\boldsymbol{\alpha}}\right)\right)\right]$ to be used in (2.1). This procedure provides some protection against nonresponse bias if $\mathbf{z}_k$ is a powerful predictor of the response probability and/or of the variable of interest (Kim and Kim 2007).

In what follows, we propose a reweighting adjustment system based on an auxiliary variable that measures the propensity of each unit to participate to the survey. To this end, further assumptions on the response model are introduced in order to assume a dependence of the p_k's on one latent auxiliary variable that is connected to the propensity scores of Rosenbaum and Rubin (1983). The proposed approach can be used when no other auxiliary information is available on $k \in s$.

3.2 Latent variables as auxiliary information

To obtain a measure of response propensities, we consider the case in which item nonresponse on the variables of interest is also present. Then, following Chambers and Skinner (2003, page 278) '*from a*

theoretical perspective the difference between unit and item nonresponse is unnecessary. Unit nonresponse is just an extreme form of item nonresponse', we assume that item response on the variables of interest is driven on respondents by the same attitude and factors that drive unit response. Latent variable models can be used to estimate such factors that, therefore, can be used as covariates in a logistic response model.

As we have already mentioned we assume that item nonresponse affects m survey variables of particular interest. A second response indicator is introduced for each item ℓ. For each item ℓ and each unit k, a binary variable $x_{k\ell}$ is defined that takes value 1 if unit k answers to item ℓ and 0 otherwise. Let $\mathbf{x}_k = (x_{k1},\ldots,x_{k\ell},\ldots,x_{km})'$ denote the vector of response indicators for unit k to the m items and let $\mathbf{y}_k = (y_{k1},\ldots,y_{k\ell},\ldots,y_{km})'$ be the study variable vector for unit k. Thus $y_{k\ell}$ is the response value of unit k to item ℓ and $x_{k\ell}$ is its response indicator.

Suppose the $x_{k\ell}$'s are related to an assumed underlying latent continuous scale; they are the indicators of a latent variable denoted by θ_k. De Menezes and Bartholomew (1996) call the variable θ_k the 'tendency to respond' to the survey. We call it here the 'will to respond to the survey' of unit k. A latent trait model with a single latent variable is used to compute θ_k for each $k \in s$ (we will see later how; see Section 4.4). Assume for the moment that θ_k is known on all sample units and, as with usual auxiliary information, can be used as a covariate. In the absence of other covariates, Model (3.3) is rewritten as

$$p_k = P(R_k = 1|\theta_k) = \frac{1}{1+\exp(-(\alpha_0+\alpha_1\theta_k))}. \tag{3.4}$$

Covariate θ_k can be viewed as a variable explaining the behavior related to the survey topic, and thus having good properties to reduce the covariance between y_{kj} and p_k and, therefore, nonresponse bias. If other suitable auxiliary information is available, it can be inserted in the model as supplementary covariates. Now, to estimate the parameters of Model (3.4), the value of θ_k has to be available for all units in the sample. The following sections provide details on how to obtain estimated values of θ_k for both respondents and nonrespondents.

4 Computing response propensities using latent trait models

The variable θ_k can be computed using a latent trait model. In general, latent variable models are multivariate regression models that link continuous or categorical responses to unobserved covariates. A latent trait model is essentially a factor analysis model for binary data (see Bartholomew, Steele, Moustaki and Galbraith 2002; Skrondal and Rabe-Hesketh 2007).

We start by creating the matrix with elements $\{x_{k\ell}\}_{k\in s;\ell=1,\ldots,m}$. Figure 4.1 shows a schematic of the indicators $x_{k\ell}$ for respondents and nonrespondents. Then, we assume that the factors that drive unit response are the same as those that drive item response on selected variables of interest. In other words, item nonresponse is assumed nonignorable.

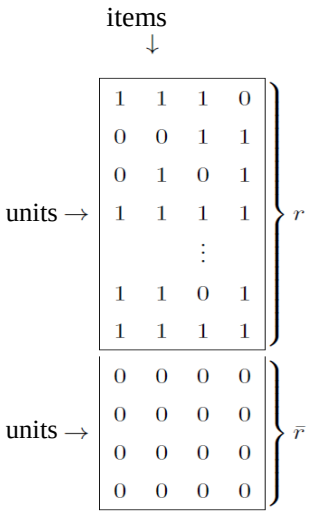


Figure 4.1 Schematic representing variables $x_{k\ell}$ for the sets r and $\bar{r}$

Let $q_{k\ell}$ be the probability of response of unit k for item ℓ, for all $\ell = 1, \ldots, m$ and $k \in r$. As in the case of unit nonresponse, $q_{k\ell}$ is modelled as a function of the variable of interest using logistic regression as follows

$$q_{k\ell} = P(x_{k\ell} = 1 | y_{k\ell}, \theta_k, R_k = 1) = \frac{1}{1 + \exp(-(\beta_{\ell 0} + \beta_{\ell 1}\theta_k + \beta_{\ell 2} y_{k\ell}))}, \tag{4.1}$$

for $\ell = 1, \ldots, m$, and $k \in r$, where $\beta_{\ell 0}$, $\beta_{\ell 1}$ and $\beta_{\ell 2}$ are parameters. Since $y_{k\ell}$ is known only for units with $x_{k\ell} = 1$, $k \in r$, Model (4.1) cannot be estimated. As in the case of unit nonresponse, we propose to estimate $q_{k\ell}$ as a function of an auxiliary variable related to the variable of interest, that is θ_k. Model (4.1) is rewritten

$$q_{k\ell} = P(x_{k\ell} = 1 | \theta_k, R_k = 1) = \frac{1}{1 + \exp(-(\beta_{\ell 0} + \beta_{\ell 1}\theta_k))}, \tag{4.2}$$

for $\ell = 1, \ldots, m$, and $k \in r$. Model (4.2) is not an ordinary logistic regression model, because the θ_k's are unobservable values taken by a latent variable. Latent trait models can be used in this case to estimate $q_{k\ell}$, θ_k and the model parameters. Note that in the area of educational testing and psychological measurement, latent trait modelling is termed Item Response Theory.

The Rasch model (Rasch 1960) is a first simple latent trait model that is well known in the psychometrical literature and used to analyze data from assessments to measure variables such as abilities and attitudes. It takes the following form

$$q_{k\ell} = \frac{1}{1 + \exp(-(\beta_{\ell 0} + \beta_1 \theta_k))} \text{ for } \ell = 1, \ldots, m \text{ and } k \in r. \tag{4.3}$$

The parameters $\beta_{\ell 0}$ are estimated for each item ℓ and reflect the extremeness (easiness) of item ℓ: larger values correspond to a larger probability of a positive response at all points in the latent space. The

parameter β_1 is known as the 'discrimination' parameter and can be fixed to some arbitrary value without affecting the likelihood as long as the scale of the individuals' propensities is allowed to be free. In many situations the assumption that item discriminations are constant across items is too restrictive. The two-parameter logistic (2PL) model generalizes the Rasch model by allowing the slopes to vary. Specifically, the 2PL model assumes the form given in Equation (4.2). The parameters $\beta_{\ell 1}$ are now estimated for each item ℓ and provide a measure of how much information an item provides about the latent variable θ_k. To achieve identifiability of Model (4.2), we can fix the value of one or more parameters $\beta_{\ell 0}$ and $\beta_{\ell 1}$ in the estimation process. Moran (1986) showed that in the 2PL model, all the parameters are identifiable under wide conditions, provided the number of items exceeds two, and all the slopes are assumed to be strictly positive. A further generalization to Model (4.2) is considered in the literature - the 3PL model - that includes another parameter, the *guessing* parameter, to model the probability that a subject with a latent variable tending to $-\infty$ responds to an item. Such an extension does not seem necessary in the context at hand and will not be considered further.

4.1 Assumptions in latent trait models

Latent trait models typically rely on the following assumptions. The first one is the so-called *conditional independence* assumption, which postulates that item responses are independent given the latent variable (i.e., the latent variable accounts for all association among the observed variables $x_{k\ell}$). Consequently, given θ_k, the conditional probability of $\mathbf{x}_k$ is

$$P(\mathbf{x}_k|\theta_k) = \prod_{\ell=1}^{m} P(x_{k\ell}|\theta_k).$$

Following Bartholomew et al. (2002, page 181) 'the assumption of conditional independence can only be tested indirectly by checking whether the model fits the data. A latent variable model is accepted as a good fit when the latent variables account for most of the association among the observed responses.'

A second assumption of Models (4.2) and (4.3) is that of *monotonicity*: as the latent variable θ_k increases, the probability of response to an item increases or stays the same across intervals of θ_k. In other words, for two values of θ_k, say a and b, and arbitrarily assuming that $a < b$, monotonicity implies that $P(x_{k\ell} = 1|\theta_k = a) < P(x_{k\ell} = 1|\theta_k = b)$ for $\ell = 1,\ldots,m$. Larger values of θ_k are associated with a greater chance of a response to each item.

Finally, the third, and possibly strongest, assumption of Models (4.2) and (4.3) is that of *unidimensionality*, implying that a single latent variable fully explains the willingness of unit k to answer the questionnaire. All these basic assumptions imply that the dependence between the items $x_{k\ell}$ may be explained by the latent variable θ_k which represents the units' willingness and that the probability that a unit k responds to a given variable increases with θ_k.

4.2 Estimation of the model

In what follows we focus on the two-parameter logistic (2PL) model given in (4.2). Let $\boldsymbol{\beta}_\ell = (\beta_{\ell 0}, \beta_{\ell 1})'$ and $\boldsymbol{\beta} = \{\boldsymbol{\beta}_\ell, \ell = 1,\ldots,m\}$. Model (4.2) can be fitted using maximum likelihood or bayesian methods. We focus here on the former. Under the maximum likelihood approach, three major

methods - joint, conditional and marginal maximum likelihood - are developed. Here, we will concentrate on marginal maximum likelihood that can be applied to fit the 2PL model. This method is also used in the simulation studies of Section 6. It consists of maximizing the likelihood of the model after the θ_k are integrated out on the basis of a common distribution assumed on these parameters. In particular, it is assumed that θ_k is a random variable following a distribution with the density function $h(\cdot)$; typically $\theta_k \sim N(0,1)$. It is also assumed that the response vectors $\mathbf{x}_k$ are independent of one another and the conditional independence assumption holds.

For a set of n_r respondents having the response vectors $\mathbf{x}_k, k = 1,\ldots,n_r$, the marginal likelihood can be expressed as

$$L\left(\boldsymbol{\beta}; \mathbf{x}_1,\ldots,\mathbf{x}_{n_r}\right) = \prod_{k=1}^{n_r} f\left(\mathbf{x}_k|\boldsymbol{\beta}\right),$$

where $f\left(\mathbf{x}_k|\boldsymbol{\beta}\right) = \int_{-\infty}^{\infty} g\left(\mathbf{x}_k|\theta_k,\boldsymbol{\beta}\right) h\left(\theta_k\right) d\theta_k$,

$$g\left(\mathbf{x}_k|\theta_k,\boldsymbol{\beta}\right) = \prod_{\ell=1}^{m} q_{k\ell}^{x_{k\ell}} \left(1-q_{k\ell}\right)^{1-x_{k\ell}} = \prod_{\ell=1}^{m} \frac{\exp\left(x_{k\ell}\left(\beta_{\ell 0}+\beta_{\ell 1}\theta_k\right)\right)}{1+\exp\left(\beta_{\ell 0}+\beta_{\ell 1}\theta_k\right)},$$

and h now denotes the density of the $N(0,1)$ distribution. The method consists in maximizing the corresponding log-likelihood, given by

$$\log L\left(\boldsymbol{\beta}; \mathbf{x}_1,\ldots,\mathbf{x}_{n_r}\right) = \sum_{k=1}^{n_r} \log\left(f\left(\mathbf{x}_k|\boldsymbol{\beta}\right)\right),$$

with respect to $\boldsymbol{\beta}$ using, for example, the EM algorithm. Estimates of $\beta_{\ell 0}$ and $\beta_{\ell 1}, \ell = 1,\ldots,m$ are thus provided. Afterwards, θ_k is estimated using the empirical Bayes method by maximizing the posterior density

$$h\left(\theta_k|\mathbf{x}_k\right) = \frac{g\left(\mathbf{x}_k|\theta_k,\boldsymbol{\beta}\right) h\left(\theta_k\right)}{g\left(\mathbf{x}_k\right)} \propto g\left(\mathbf{x}_k|\theta_k,\boldsymbol{\beta}\right) h\left(\theta_k\right),$$

with respect to θ_k and keeping item parameters and observations fixed. Estimates of $q_{k\ell}$ are obtained using Expression (4.2), where $\beta_{\ell 0}, \beta_{\ell 1}$ and θ_k are replaced with their estimates.

4.3 Goodness-of-fit measures of the model

Different goodness-of-fit measures are proposed in the literature to test whether the model given in (4.2) adequately fits the data (see e.g., Bartholomew et al. 2002). One uses two-way and three-way margins of the response items. Discrepancies between the expected (E) and observed (O) counts in these tables are measured using the statistic $R = (O-E)^2/E$. Large values of R for the second-order or third-order margins will identify sets of items for which the model does not fit well. Note that the residuals $(O-E)^2/E$ are not independent and they cannot be summed to give an overall test statistics

distributed as a chi-squared (see Bartholomew et al. 2002, page 186). Item fit indexes (Bond and Fox 2007) can be used to this end as well. On the basis of estimated latent variables and item parameters, the expected response of a unit to an item can be computed. The similarity between the observed and expected responses to any item can be assessed through two fit mean-square statistics: the outlier-sensitive fit statistic (item outfit) and the information-weighted fit statistic (item infit). The estimate produced by the item outfit is relatively more affected by unexpected responses different from a person's measure, i.e., it is more sensitive to unexpected observations by units on items that are relatively very easy or very hard for them to answer. The item infit has each observation weighted by the information and, on the other side, is relatively more affected by unexpected responses closer to a person's measure, i.e., it is more sensitive to unexpected patterns of observations by units on items that are roughly targeted on them according to their latent variable value. The expected value for both statistics is one. For infit and outfit values greater/less than one indicate more/less variation between the observed and the predicted response patterns, a range of 0.5 to 1.5 is generally acceptable (Bond and Fox 2007).

In addition, point-measure correlations (Olsson, Drasgow and Dorans 1982) can be used to estimate the correlation between the latent variable and the single item response. Items for which such measures take negative or zero values should be removed from the analysis or may be evidence that the latent construct is not unidimensional. Unidimensionality can be tested by running a Principal Components Analysis (PCA) of the standardized residuals for the items (Wright 1996). In this way the first component (dimension) has already been removed, and it is possible to look at secondary dimensions, components or contrasts. Unidimensionality is supported by observing that the eigenvalue of the first PCA component in the correlation matrix of the residuals is small (usually less than 2.0). If not, the loadings on the first contrast indicate that there are contrasting patterns in the residuals.

Finally, when items are used to form a scale, they need to have internal consistency. Cronbach alpha can be used to test whether items have the reliability property, i.e., if they all measure the same thing, then they should be correlated with one another.

4.4 Estimation of p_k

Two solutions are shown here to estimate p_k using information from the latent trait model. The first solution uses logistic regression to estimate p_k for all $k \in s$, and a two-stage approach.

Stage 1: First, an estimate $\hat{\theta}_k$ of θ_k is provided. To compute a value $\hat{\theta}_k$ for $k \in \bar{r}$, we assume again that unit nonresponse is just an extreme form of item nonresponse. Thus, a nonrespondent does not answer any item ℓ and thus $x_{k\ell} = 0$, for all $\ell = 1, \ldots, m$. The computation of $\hat{\theta}_k$ for $k \in \bar{r}$ is handled as follows: we add to the set r a phantom respondent unit $\tilde{k}$ having $x_{\tilde{k}\ell}$ equal to 0, for all $\ell = 1, \ldots, m$. We denote this new set by $\tilde{r} = r \cup \{\tilde{k}\}$. We estimate the parameters of Model (4.2) using all units $k \in \tilde{r}$, and compute the values $\hat{\theta}_k, k \in \tilde{r}$. Model (4.2) allows the computation of $\hat{\theta}_k$ for all $k \in \tilde{r}$. Unit $\tilde{k}$ has an estimated value $\hat{\theta}_{\tilde{k}}$. We assign to all units $k \in \bar{r}$ an estimate $\hat{\theta}_k$ equal to $\hat{\theta}_{\tilde{k}}$. Thus, the same value of $\hat{\theta}_k$ is provided for all $k \in \bar{r}$. Using this method, each unit $k \in s$ has associated an estimate $\hat{\theta}_k$. This is the key feature for the estimation of the response probabilities p_k provided in the next stage.

Stage 2: We use the estimate $\hat{\theta}_k$, for $k \in s$, provided in the first stage as a covariate in Model (3.4) instead of the unknown value of θ_k; in particular

$$p_k = P\left(R_k = 1 \middle| \hat{\theta}_k\right) = \frac{1}{1 + \exp\left(-\left(\alpha_0 + \alpha_1 \hat{\theta}_k\right)\right)}, \text{for all } k \in s. \tag{4.4}$$

Model (4.4) provides estimates $\hat{p}_k$ of p_k, for all $k \in s$.

One of the Referees suggested the following solution to estimate p_k. Let $S_k = \sum_{\ell=1}^{m} x_{k\ell}$ be the raw score for unit k, i.e., the number of items unit k has responded to: if $k \in \bar{r}$, then $S_k = 0$; if $k \in r$, then $S_k > 0$. Then p_k can be estimated by modelling $P(S_k > 0|\theta_k)$. By the conditional independence assumption we have

$$\begin{aligned} p_k &= P(S_k > 0|\theta_k) = 1 - P(S_k = 0|\theta_k) = 1 - P\left(\bigcap_{\ell=1}^{m}(x_{k\ell} = 0|\theta_k)\right) \\ &= 1 - \prod_{\ell=1}^{m}(1 - P(x_{k\ell} = 1|\theta_k)). \end{aligned}$$

We have $P(x_{k\ell} = 1|\theta_k) = P(R_k = 1|\theta_k)P(x_{k\ell} = 1|\theta_k, R_k = 1) + P(R_k = 0|\theta_k)P(x_{k\ell} = 1|\theta_k, R_k = 0) = p_k q_{k\ell}$, because $P(x_{k\ell} = 1|\theta_k, R_k = 0) = 0$. As a result, we obtain

$$p_k = 1 - \prod_{\ell=1}^{m}(1 - p_k q_{k\ell}), k \in r.$$

The estimated response probability $\hat{p}_k$, $k \in r$ is obtained as a solution to the polynomial equation

$$\hat{p}_k = 1 - \prod_{\ell=1}^{m}(1 - \hat{p}_k \hat{q}_{k\ell}).$$

This solution, although very elegant, has two drawbacks. If m is large, the above polynomial equation is difficult or even impossible to solve. If it possible to solve the polynomial equation for moderate m, the real solutions are not necessarily in (0, 1). This solution has not been considered here further.

5 The proposed estimator and its variance estimation

Recall that we have a variable of particular interest y_j and that item nonresponse is present for it. If we wish to estimate the population total Y_j of y_j, then a naive estimator that does not correct neither for unit nor for item nonresponse is given by

$$\hat{Y}_{j,\text{naive}} = N \sum_{k \in r_j} \frac{y_{kj}}{\pi_k} \Big/ \sum_{k \in r_j} \frac{1}{\pi_k}. \tag{5.1}$$

Reweighting item responders is also an approach to handle item nonresponse. Moustaki and Knott (2000) propose to weight item responders by the inverse of the fitted probability of item response $\hat{q}_{k\ell}$,

assuming $\hat{q}_{k\ell} > 0$. Therefore, a possible adjustment weight for item and unit nonresponse associated with unit $k \in r_j$ is given by $1/\left(\hat{p}_k \hat{q}_{kj}\right)$. We propose using the three-phase estimator adjusted for item and unit nonresponse via reweighting given by

$$\hat{Y}_{j,pq} = \sum_{k \in r_j} \frac{y_{kj}}{\pi_k \hat{p}_k \hat{q}_{kj}}, \tag{5.2}$$

where $\hat{p}_k$ is provided by Model (4.4), and $\hat{q}_{kj}$ by Model (4.2). Proposals that use imputation of y_{kj} values for $k \in r \setminus r_j$ to deal with item nonresponse are also considered but not reported for reasons of space. They are available from the Authors upon request.

The properties of the proposed estimator (5.2) depend on the assumptions made about the unit and the item nonresponse mechanisms. In particular, Estimator (5.2) assumes a second phase of sampling with unknown response probabilities. If we ignore estimation of θ_k in Model (4.4), the results in Kim and Kim (2007) on design consistency of the two-phase estimator that uses estimated response probabilities hold here as well when considering maximum likelihood estimates for the parameters α_0 and α_1. Again, ignoring estimation of the latent variable θ_k and using marginal maximum likelihood estimates for the parameters $\beta_{\ell 0}$ and $\beta_{\ell 1}$ in Model (4.2), estimator $\hat{Y}_{j,pq}$ will be consistent if the models for unit and item nonresponse probabilities are correctly specified.

We can consider replication methods for variance estimation of the proposed estimator and combine proposals for two-phase sampling (Kim, Navarro and Fuller 2006) and for generalized calibration in the presence of nonresponse (Kott 2006). In particular, the replicate variance estimator can be written as

$$\hat{V}_r = \sum_{l=1}^{L} c_l \left(\hat{Y}_{j,pq}^{(l)} - \hat{Y}_{j,pq}\right)^2,$$

where $\hat{Y}_{j,pq}^{(l)}$ is the l^{th} version of $\hat{Y}_{j,pq}$ based on the observations included in the l^{th} replicate, L is the number of replications, c_l is a factor associated with replicate l determined by the replication method. The l^{th} replicate of $\hat{Y}_{j,pq}$ can be written as $\hat{Y}_{j,pq}^{(l)} = \sum_{k \in r_j} w_{3k}^{(l)} y_{kj}$, where $w_{3k}^{(l)}$ denotes the replicate weight for the k^{th} unit in the l^{th} replication. These replicate weights are computed using a two-step procedure.

First, note that, if we ignore for the moment the presence of item nonresponse, the two-phase estimator $\hat{Y}_{j,p} = \sum_{k \in r} w_{2k} y_{kj}$, has weights

$$w_{2k} = 1/(\pi_k p_k) = w_{1k} F\left(\hat{\theta}_k; \alpha_0, \alpha_1\right),$$

with, $w_{1k} = 1/\pi_k$, $F\left(\hat{\theta}_k; \alpha_0, \alpha_1\right) = 1 + \exp\left(-\left(\alpha_0 + \alpha_1 \hat{\theta}_k\right)\right)$ (see Equation (4.4)). Let $\hat{\mathbf{z}}_1 = \sum_{k \in s} w_{1k} \mathbf{z}_{1k}$ be the first phase estimate of the total of variable $\mathbf{z}_1$ defined as $\mathbf{z}_{1k} = \pi_k p_k \left(1, \hat{\theta}_k\right)'$. Then, parameters α_0 and α_1 are such that

$$\sum_{k \in r} w_{1k} F\left(\hat{\theta}_k; \alpha_0, \alpha_1\right) \mathbf{z}_{1k} = \hat{\mathbf{z}}_1. \tag{5.3}$$

This procedure is equivalent to obtaining unweighted maximum likelihood estimates, but is convenient to set it as a non-linear generalized calibration problem. In this way, it is possible to use the approach in Kott (2006), combined with that in Kim et al. (2006), to obtain replicate weights using the following steps.

Step 1: Compute the first phase estimate of the total of $\mathbf{z}_{1k}$ with l^{th} observation deleted, i.e., $\hat{\mathbf{z}}_1^{(l)} = \sum_{k \in s} w_{1k}^{(l)} \mathbf{z}_{1k}$, where $w_{1k}^{(l)}$ is the classical jackknife replication weight for unit k in replication l. Compute the jackknife weights for the second phase sampling using $\hat{\mathbf{z}}_1^{(l)}$ as a benchmark. In particular, $w_{2k}^{(l)}$ are chosen to be $w_{2k}^{(l)} = w_{2k} w_{1k}^{(l)} F\left(\hat{\theta}_k; \alpha_0, \alpha_1\right) / w_{1k}$ with α_0 and α_1 such that

$$\sum_{k \in r} w_{2k}^{(l)} \mathbf{z}_{1k} = \hat{\mathbf{z}}_1^{(l)}.$$

This procedure provides weights that are very similar to those considered in Kott (2006) and can be computed using existing software that handles generalized calibration.

Item nonresponse is handled similarly by considering $w_{3k} = 1/\left(\pi_k p_k q_{kj}\right) = w_{2k} F\left(\hat{\theta}_k; \beta_{j0}, \beta_{j1}\right)$ (compare Equation (4.3)). A major approximation here is to assume that, given $\hat{\theta}_k$, parameters β_{j0} and β_{j1} are estimated using a classical logistic model (instead of a 2PL model) and are such that

$$\sum_{k \in r_j} w_{2k} F\left(\hat{\theta}_k; \beta_{j0}, \beta_{j1}\right) \mathbf{z}_{2k} = \hat{\mathbf{z}}_2,$$

where $\hat{\mathbf{z}}_2 = \sum_{k \in r} w_{2k} \mathbf{z}_{2k}$ and $\mathbf{z}_{2k} = \pi_k p_k q_{kj} \left(1, \hat{\theta}_k\right)^T$. Another drawback is that auxiliary variables $\mathbf{z}_{2k}$ depend on j and, therefore, different sets of weights have to be produced for the different variables of interest.

Step 2: Third phase jackknife weights are obtained by first computing the second phase estimate of the total of $\mathbf{z}_{2k}$ with unit l removed by using weights coming from Step 1, i.e., $\hat{\mathbf{z}}_2^{(l)} = \sum_{k \in r} w_{2k}^{(l)} \mathbf{z}_{2k}$. Then, using $\hat{\mathbf{z}}_2^{(l)}$ as a benchmark, $w_{3k}^{(l)}$ are chosen to be $w_{3k}^{(l)} = w_{3k} w_{2k}^{(l)} F\left(\hat{\theta}_k; \beta_{j0}, \beta_{j1}\right) / w_{2k}$ with β_{j0} and β_{j1} computed via

$$\sum_{k \in r_j} w_{3k}^{(l)} \mathbf{z}_{2k} = \hat{\mathbf{z}}_2^{(l)}.$$

6 Simulation studies

We evaluate the performance of the estimator presented in Section 5 by means of a Monte Carlo simulation under two different settings. The first one uses a real data set as the population and considers variables of interest that are all binary, while the second one uses simulated population data with variables of interest that are continuous. Results from the first setting are presented in Section 6.1, while those from the second setting are presented in Section 6.2.

In both settings, simple random sampling without replacement is employed and the following estimators are considered:

- $\mathrm{HT} = \sum_{k \in s} y_{kj} / \pi_k$: the Horvitz-Thompson estimator in the case of full response is computed as a benchmark in the absence of nonresponse.

- $\hat{Y}_{j,\mathrm{naive}}$: the naive estimator given in (5.1); no explicit action is taken to adjust for unit and item nonresponse. Note that for simple random sampling without replacement, it reduces to $\hat{Y}_{j,\mathrm{naive}} = N \sum_{k \in r_j} y_{kj} / n_{r_j}$, where n_{r_j} is the size of the set r_j, and it is the same as the Horvitz-Thompson estimator adjusted for unit nonresponse that assumes uniform response probabilities estimated by n_{r_j} / n.

- $\hat{Y}_{j,pq}$: the three-phase estimator proposed in Section 5, Equation (5.2).

- $\hat{Y}_{j,pq,\mathrm{true}}$: the three-phase estimator that uses the true values for the response probabilities p_k and q_{kj} is also computed for comparison with $\hat{Y}_{j,pq}$ to understand the effect of estimating the response probabilities.

The simulations are carried out in R version 2.15, using the R package '`ltm`' (Rizopoulos 2006) to fit the latent trait models. The following performance measures are computed for each estimator, generically denoted below by $\hat{Y}$ where suffix j is dropped for ease of notation (Y denotes the population total):

- the Monte Carlo Bias

$$\mathrm{B} = E_{\mathrm{sim}}\left(\hat{Y}\right) - Y,$$

where $E_{\mathrm{sim}}\left(\hat{Y}\right) = \sum_{i=1}^{M} \hat{Y}_i / M$, $\hat{Y}_i$ is the value of the estimator $\hat{Y}$ at the i^{th} simulation run and M is total number of simulation runs;

- the Relative Bias

$$\mathrm{RB} = \frac{\mathrm{B}}{Y};$$

- the Monte Carlo Standard Deviation

$$\sqrt{\mathrm{VAR}} = \sqrt{\frac{1}{M-1} \sum_{i=1}^{M} \left(\hat{Y}_i - E_{\mathrm{sim}}\left(\hat{Y}\right)\right)^2};$$

- the Monte Carlo Mean Squared Error

$$\mathrm{MSE} = \mathrm{B}^2 + \mathrm{VAR}.$$

6.1 Simulation setting 1

We consider the Abortion data set formed by four binary variables extracted from the 1986 British Social Attitudes Survey and concerning the attitude towards abortion. The data is available in the R

package '`ltm`' (Rizopoulos 2006). $N = 379$ individuals answered the following questions after being asked if the law should allow abortion under the circumstances presented under each item:

1. The woman decides on her own that she does not wish to keep the baby.
2. The couple agrees that they do not wish to have a child.
3. The woman is not married and does not wish to marry the man.
4. The couple cannot afford any more children.

The variable of interest y_j is selected to be the second one $(j = 2)$ with a total $Y_j = 225$ in the population.

The data is analyzed by Bartholomew et al. (2002) as an example in which a latent variable can be found that measures the attitude towards abortion. At the population level, we compute the latent variable (denoted here by θ_k^a) using Model (4.2) on the $\{y_{k\ell}\}_{k=1,\ldots,N;\ell=1,\ldots,4}$ data. The correlation between the values $y_{k\ell}$ and θ_k^a is approximatively equal to 0.85, for $\ell = 1,\ldots,4$. Afterwards, we have set $\theta_k = \hat{\theta}_k^a$, for all $k = 1,\ldots,N$.

At the population level, the unit response probabilities are generated using the following response model

$$p_k = 1/(1 + \exp(-(0.7 + y_{k2} + \theta_k + 0.2\varepsilon_k))), \tag{6.1}$$

with $\varepsilon_k \sim U(0,1)$, to simulate nonignorable nonresponse. The population mean of p_k is approximately 0.74.

To generate item response probabilities at the population level, the following model is used

$$q_{k\ell} = 1/(1 + \exp(-(b_\ell\theta_k + a_\ell + y_{k\ell}))), \quad \text{for } \ell = 1,\ldots,4, \tag{6.2}$$

where $b_\ell = 3$, for $\ell = 1,\ldots,4$, while a_ℓ takes different values according to ℓ; in particular, $a_1 = 1, a_2 = 0, a_3 = -0.5$ and $a_4 = 1$. The nominal item nonresponse rate for the four items in the population is 35%, 42%, 47%, 31%, respectively.

We draw $M = 10{,}000$ simple random samples without replacement from the population using two sample sizes: $n = 50$ and $n = 100$. In each sample s, the units are classified as respondents according to Poisson sampling, using the probabilities p_k computed as in Equation (6.1) and resulting in the set r. Then, given r, the matrix $\{x_{k\ell}\}_{k\in r;\ell=1,\ldots,4}$ is constructed where the values $x_{k\ell}$ are drawn using Poisson sampling with probabilities $q_{k\ell}$ defined in (6.2). In each simulation run, Model (4.2) and the respondents set r are used to compute the variable $\hat{\theta}_k$ for all $k \in s$ as described in Section 4.4. Model (4.4) is fitted to obtain $\hat{p}_k$. The average item nonresponse rate over simulations for the four items is found to be 26%, 33%, 38% and 23%. The jackknife variance estimator was computed as described in Section 5 using the `gencalib()` function in R package 'sampling' (Tillé and Matei 2012) and the logistic distance (Deville, Särndal and Sautory 1993).

Table 6.1 reports the results for $n = 50$ and $n = 100$. As expected, HT and $\hat{Y}_{j,pq,\text{true}}$ have almost zero bias, with the second one showing a relatively larger MSE that is due uniquely to the smaller sample

size. The naive estimator shows a very large negative bias. This is due to the fact that units with a zero value of y_j are less likely to respond and the total is clearly underestimated. The estimator $\hat{Y}_{j,pq}$ shows a much smaller bias than the naive estimator. Note that the performance of the proposed estimator is mostly driven by absolute bias, so that the performance is not particularly different when increasing the sample size, apart from a decrease in variance. If we compare $\hat{Y}_{j,pq,\text{true}}$ and $\hat{Y}_{j,pq}$, we note that $\hat{Y}_{j,pq}$ still suffers from some bias that comes from response model misspecification (we are not accounting for the variables of interest values).

For the proposed estimator, the jackknife variance estimator was also tested by looking at the empirical coverage of a 95% confidence interval computed for each replicate as $\hat{Y}_{j,pq} \pm 1.96\sqrt{\hat{V}_r}$. For $n = 50$, the mean value of $\sqrt{\hat{V}_r}$ over simulations was 54.8, while for $n = 100$, 53.3, with a 95% coverage rate of 94.6% and 96.3%, respectively. The replicate estimator overestimates the Monte Carlo standard deviation reported for $\hat{Y}_{j,pq}$ in Table 6.1 in both cases, but shows good coverage rates.

Table 6.1
Simulation results for setting 1 - Abortion data set

Estimator	B	$\sqrt{\text{VAR}}$	MSE	% RB
$n = 50$				
HT	0.05	24.5	600.5	< 0.1
$\hat{Y}_{j,\text{naive}}$	-126.5	19.4	16,378.6	-56.2
$\hat{Y}_{j,pq}$	20.6	32.4	1,474.1	9.1
$\hat{Y}_{j,pq,\text{true}}$	0.02	35.0	1,225.0	< 0.1
$n = 100$				
HT	-0.06	16.0	255.5	< 0.1
$\hat{Y}_{j,\text{naive}}$	-126.9	13.5	16,284.1	-56.4
$\hat{Y}_{j,pq}$	17.9	21.9	802.2	8.0
$\hat{Y}_{j,pq,\text{true}}$	-0.1	23.7	559.9	< 0.1

To study the performance of the latent model on the population level and the correlation between the variable of interest and the estimated latent variable, we apply the procedure described earlier using $q_{k\ell}$ defined in (6.2) to construct the matrix $\{x_{k\ell}\}_{k=1,\ldots,N;\ell=1,\ldots,4}$ for all population units. We fit Model (4.2) on the population level and compute the variable θ_k for all $k = 1,\ldots,N$. The Cronbach's alpha measure takes value 0.83 showing a good internal consistency of the items. The correlation coefficient between the variable of interest and the estimated latent variable takes value 0.76, indicating that the latent auxiliary information has a strong power of predicting y_{k2}, as advocated in the model of Cassel et al. (1983). Inspection of the two-way margins for the matrix $\{x_{k\ell}\}$ gives the residuals $(O-E)^2/E$ between 0.03 and 0.23. Similarly, the three-way margins for the matrix $\{x_{k\ell}\}$ give residuals between 0 and 1.19. This indicates that we have no reason to reject here the one-factor latent Model (4.2) (see Bartholomew et al. 2002, page 186).

6.2 Simulation setting 2

We generate $\{y_{k1},\ldots,y_{k6},\theta_k\}$ for $k = 1,\ldots, N = 2{,}000$ using a multivariate normal distribution with mean 1. The degree of correlation between y_ℓ and $y_{\ell'}$ is 0.8, with $\ell,\ \ell' = 1,\ldots,6,\ \ell \neq \ell'$. We set the variable of interest to be y_6 and consider different degrees of correlation between its values and those taken by θ_k, namely 0.3, 0.5, 0.8. The values of θ_k are afterwards standardized to have mean 0 and variance 1.

The response probabilities are obtained by first computing

$$p_k^\circ = 1/[1+\exp(-(0.5+y_{k1}+\theta_k))], \quad \text{for } k = 1,\ldots,N, \tag{6.3}$$

and then rescaling them to take values between 0.1 and 0.9, with a population mean approximatively equal to 0.7.

The item response probabilities are generated by first computing

$$q_{k\ell}^\circ = 1/(1+\exp(-(b_\ell\theta_k + a_\ell + y_{k\ell}))), \quad \text{for } k = 1,\ldots,N \text{ and } \ell = 1,\ldots,6, \tag{6.4}$$

where $\{a_\ell\}_{\ell=1,\ldots,6} = \{1,0,-0.5,1,0,-0.5\}$ and $\{b_\ell\}_{\ell=1,\ldots,6} = \{1,1,1,1.5,1.5,1.5\}$, and then rescaling the values to be between 0.1 and 0.95.

We draw $M = 10{,}000$ samples by simple random sampling without replacement of size $n = 200$. For each sample s, a response set r is created by carrying out Poisson sampling with parameter p_k defined in (6.3). Each element of the matrix $\{x_{k\ell}\}_{k\in r,\ell=1,\ldots,6}$ is generated using Poisson sampling with parameter $q_{k\ell}$ defined in (6.4). Item nonresponse rates over simulations take approximately value 18%, 28%, 35%, 19%, 29%, 34%, for $\ell = 1,\ldots,6$, respectively. For each simulation run, Model (4.2) is used to compute the variable $\hat{\theta}_k$ for all $k \in s$. Model (4.4) is then fitted to obtain $\hat{p}_k$.

Table 6.2
Simulation results for setting 2 - Simulated continuous data

Estimator	B	$\sqrt{\text{VAR}}$	MSE	RB%
correlation coefficient 0.3				
HT	-0.7	131.6	17,331.2	≈ -0.0
$\hat{Y}_{j,\text{naive}}$	825.6	177.1	713,039.3	41.0
$\hat{Y}_{j,pq}$	-227.4	188.0	87,033.0	-11.3
$\hat{Y}_{j,pq,\text{true}}$	48.4	231.8	56,073.2	2.4
correlation coefficient 0.5				
HT	0.1	135.0	18,220.5	≈ 0.0
$\hat{Y}_{j,\text{naive}}$	972.6	176.2	977,009.5	50.7
$\hat{Y}_{j,pq}$	-180.0	175.5	63,552.0	-9.4
$\hat{Y}_{j,pq,\text{true}}$	74.8	212.7	50,844.0	3.9
correlation coefficient 0.8				
HT	-0.1	134.1	17,992.0	≈ -0.0
$\hat{Y}_{j,\text{naive}}$	1,154.6	168.1	1,361,388.1	57.7
$\hat{Y}_{j,pq}$	-184.8	164.4	61,173.0	-9.2
$\hat{Y}_{j,pq,\text{true}}$	100.6	196.2	48,597.9	5.0

Table 6.2 reports on the performance of the estimators for the three values taken by the nominal correlation coefficient between y_{k1} and θ_k : $0.3, 0.5$ and 0.8. The proposed estimator is always able to reduce bias over the naive estimator, even when the correlation between the variable of interest and the latent variable gets smaller. The relative bias takes acceptable values in most cases. Bias deserves a closer look. The naive estimator in all cases largely overestimates the total. This is expected, because the values p_k, q_{k6}, θ_k and y_{k6} all go in the same direction. Therefore, in our respondents sample, we are more likely to find relative larger values for y_6 by this providing overestimation for the naive estimator. On the other hand, $\hat{Y}_{j,pq}$ underestimates the total because it is based only on the observed units of r_j that do have relatively large values for y_6, but also relatively large values for p_k and q_{k6} and, therefore, end up having a small weight.

The matrix of population values $\{x_{k\ell}\}_{k=1;..;2,000;\ell=1,..,6}$ is constructed in the same way as in Section 6.1 to validate the assumptions behind the 2PL model. The Cronbach's alpha takes approximately value 0.5 for the correlation coefficient equal to 0.3, 0.6 for 0.5, and 0.7 for 0.8; the pairwise association between the six items reveals p-values smaller than 0.01. Inspection of the two-way and three-way margins of the matrix $\{x_{k\ell}\}$ gives residuals $(O-E)^2/E$ that all take values smaller than 4. Therefore, the one factor latent model can be accepted and items all seem to be measuring the same latent trait.

7 Discussion and conclusions

We have proposed a reweighting system to compensate for non-ignorable nonresponse based on a latent auxiliary variable. This variable is computed for each unit in the sample using a latent model assuming the existence of item nonresponse and that the same latent structure is hidden behind item and unit nonresponse. Unit response probabilities are then estimated by a logistic model that uses as a covariate the latent trait extracted by the response patterns using a latent trait model. The proposed reweighting system is then used in a three-phase estimator to handle nonresponse, together with a replication method to estimate its uncertainty. The main goal is to reduce nonresponse bias in the estimation of the population total. The proposed estimator performs well in our simulation studies compared with the naive estimator, and the gain in efficiency is substantial in certain cases. Reductions in bias are also seen when the correlation between the latent trait and the variable of interest is modest.

By design, the estimated latent variable $\hat{\theta}_k$ is related to the response indicators x_{kj} for the variable of interest y_j; since nonresponse is assumed to be non-ignorable, y_{kj} and x_{kj} are related as well. If the following condition holds,

$$\rho^2_{y_j,x_j} + \rho^2_{\hat{\theta},x_j} > 1,$$

where the correlation coefficients $\rho_{y_j,x_j}, \rho_{\hat{\theta},x_j} > 0$, then y_j and $\hat{\theta}$ are positively correlated (see Langford, Schwertman and Owens 2001). Note that the minimum degree of correlation between the variable of interest and the latent variable capable of reducing the nonresponse bias was found to be 0.3 in simulation setting 2 (Section 6.2). Of course, bias reduction depends on model assumptions. If response indicators are not good predictors of unit response behavior, then model misspecification is present and, of course, reduction in bias may not be present and variance could be introduced in estimation. Nonetheless,

diagnostic tools from item response theory can be used to assess the goodness of fit of the latent trait model employed to estimate values for θ_k.

We have considered the case in which no auxiliary information is available at the sample or population level to reduce nonresponse bias. Observed covariates (if available) and the latent variable can be, however, used together in the estimation of response probabilities. Moreover, latent trait models can, themselves, be fitted with covariates. The introduction of covariates in these models should be carried out with increasing prudence on variance.

The proposed estimator is a three-phase estimator using a reweighting system based on $\hat{p}_k$ and $\hat{q}_{kj}$. It is known that small values of $\hat{p}_k$ and $\hat{q}_{kj}$ may lead to unstable reweighted estimators because of large nonresponse weights. To overcome this problem, the propensity score method (e.g., Eltinge and Yansaneh 1997) is often used in practice, providing a good solution against extreme weights adjustments. In order to apply this method in our framework, the respondents to y_j should be grouped in different classes given by the quantiles of $1/\left(\hat{p}_k \hat{q}_{kj}\right)$. The final step is the calculation of a weight for each class.

Final remarks concern the conditional independence assumption in latent trait models. In nonresponse literature, it is usual to use Poisson sampling to model unit response behavior by assuming that units in the set r are selected with unknown response probabilities and that response is independent from unit to unit. The conditional independence assumption in the latent trait models is a similar condition applied to items. Both assumptions are strong, sometimes they are in doubt, yet they are necessary in the statistical inferential process.

Different methods were developed in psychometric literature to relax the conditional independence assumption. We cite here the *partial independence* approach by Reardon and Raudenbush (2006), developed for the case where responses to earlier questions determine whether later questions are asked or not, and where the usual conditional independence assumption of standard models fails. This approach could be used in our framework for the case where $q_{k\ell}$ is defined as $P(x_{k\ell} = 1|x_{kj},$ for some $j \in \{1,\ldots,m\}, \ell \neq j, \theta_k)$ instead of $P(x_{k\ell} = 1|\theta_k), k \in r.$ Another useful approach for cases where items are clustered is the latent trait hierarchical modeling. A random effect is introduced into a latent trait model to account for potential residual dependence due to the common sources of variation shared by clusters of items (see e.g., Scott and Ip 2002). Further research should be done to accommodate these approaches in the survey sampling framework.

Acknowledgements

The work of M. Giovanna Ranalli has been developed partially under the support of the project PRIN-SURWEY (grant 2012F42NS8, Italy).

References

Bartholomew, D.J., Steele, F., Moustaki, I. and Galbraith, J.I. (2002). *The Analysis and Interpretation of Multivariate Data for Social Scientists*. Chapman and Hall/CRC.

Beaumont, J.-F. (2000). An estimation method for nonignorable nonresponse. *Survey Methodology,* 26, 2, 131-136.

Bethlehem, J. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics,* 4, 3, 251-260.

Biemer, P.P., and Link, M.W. (2007). *Evaluating and Modeling early Cooperator Effects in RDD Surveys*. New York: John Wiley & Sons, Inc.

Bond, T., and Fox, C. (2007). *Applying the Rasch model: Fundamental Measurement in the Human Sciences* (2nd Ed.). Lawrence Erlbaum Associates, Inc, Mahwah, N.J.

Cassel, C.M., Särndal, C.-E. and Wretman, J.H. (1983). Some uses of statistical models in connection with the nonresponse problem. In *Incomplete Data in Sample Surveys,* (Eds., W.G. Madow and I. Olkin), New York: Academic Press. 3, 143-160.

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

Copas, A.J., and Farewell, V.T. (1998). Dealing with non-ignorable non-response by using an 'enthusiasm-to-respond' variable. *Journal of the Royal Statistical Society, Series A,* 161, 385-396.

De Menezes, L.M., and Bartholomew, D.J. (1996). New developments in latent structure analysis applied to social attitudes. *Journal of Royal Statistical Society, Series A,* 159, 213-224.

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

Drew, J.H., and Fuller, W.A. (1980). Modeling nonresponse in surveys with callbacks. *Proceedings of the Survey Research Methods Section,* American Statistical Association.

Eltinge, J.L., and Yansaneh, I.S. (1997). Diagnostics for formation of nonresponse adjustment cells, with an application to income nonresponse in the U.S. Consumer Expenditure Survey. *Survey Methodology,* 23, 1, 33-40.

Greenlees, J.S., Reece, W.S. and Zieschang, K.D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association,* 77, 251-261.

Groves, R.M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly,* 70, 5, 646-675.

Groves, R.M., Couper, M., Presser, S., Singer, E., Tourangeau, R., Acosta, G.P. and Nelson, L. (2006). Experiments in producing nonresponse bias. *Public Opinion Quarterly,* 70, 5, 720-736.

Kim, J.K., and Kim, J.J. (2007). Nonresponse weighting adjustment using estimated response probability. *Canadian Journal of Statistics,* 35, 501-514.

Kim, J.K., Navarro, A. and Fuller, W.A. (2006). Replication variance estimation for two-phase stratified sampling. *Journal of the American Statistical Association,* 101, 473, 312-320.

Kott, P.S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology,* 32, 2, 133-142.

Langford, E., Schwertman, N. and Owens, M. (2001). Is the property of being positively correlated transitive? *The American Statistician,* 55, 4, 322-325.

Legg, J.C., and Fuller, W.A. (2009). Two-phase sampling. *Handbook of Statistics,* 29, 55-70.

Little, R.J., and Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey means? *Survey Methodology,* 31, 2, 161-168.

Little, R.J.A., and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons, Inc.

Moran, P.A.P. (1986). Identification problems in latent trait models. *British Journal of Mathematical and Statistical Psychology,* 39, 2, 208-212.

Moustaki, I., and Knott, M. (2000). Weighting for item non-response in attitude scales using latent variable models with covariates. *Journal of Royal Statistical Society, Series A,* 163, 445-459.

Oh, H.L., and Scheuren, F.J. (1983). Weighting adjustments for unit non-response. In *Incomplete Data in Sample Surveys,* (Eds., W.G. Madow, I. Olkin and D.B. Rubin). New York: Academic Press. 2, 143-184.

Olsson, U., Drasgow, F. and Dorans, N. (1982). The polyserial correlation coefficient. *Psychometrika,* 47, 337-347.

Qin, J., Leung, D. and Shao, J. (2002). Estimation with survey data under nonignorable nonresponse or informative sampling. *Journal of the American Statistical Association,* 97, 193-200.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. *The Danish Institute of Educational Research,* Copenhagen.

Reardon, S.F., and Raudenbush, S.W. (2006). A partial independence item response model for surveys with filter questions. *Sociological Methodology,* 36, 1, 257-300.

Rizopoulos, D. (2006). `ltm`: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software,* 17, 5, 1-25.

Rosenbaum, P.R., and Rubin, D.B. (1983). The central role of the propensity score inobservational studies for causal effects. *Biometrika,* 70, 41-55.

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

Scott, S.L., and Ip, E.H. (2002). Empirical bayes and item-clustering effects in a latent variable hierarchical model: A case study from the national assessment of educational progress. *Journal of American Statistical Association,* 97, 459, 1-11.

Skrondal, A., and Rabe-Hesketh, S. (2007). Latent variable modelling: A survey. *Scandinavian Journal of Statistics,* 34, 712-745.

Tillé, Y., and Matei, A. (2012). *Sampling: Survey Sampling*. R package version 2.5.

Wright, B. (1996). Local dependency, correlations and principal components. *Rasch Meas Trans,* 10-3, 509-511.

Zhang, L.C. (2002). A method of weighting adjustment for survey data subject to nonignorable nonresponse. DACSEIS research paper no. 2, http://w210.ub.unituebingen.de/dbt/volltexte/2002/451.

Survey Methodology, June 2015
Vol. 41, No. 1, pp. 165-181
Statistics Canada, Catalogue No. 12-001-X

One step or two? Calibration weighting from a complete list frame with nonresponse

Phillip S. Kott and Dan Liao[1]

Abstract

When a random sample drawn from a complete list frame suffers from unit nonresponse, calibration weighting to population totals can be used to remove nonresponse bias under either an assumed response (selection) or an assumed prediction (outcome) model. Calibration weighting in this way can not only provide double protection against nonresponse bias, it can also decrease variance. By employing a simple trick one can estimate the variance under the assumed prediction model and the mean squared error under the combination of an assumed response model and the probability-sampling mechanism simultaneously. Unfortunately, there is a practical limitation on what response model can be assumed when design weights are calibrated to population totals in a single step. In particular, the choice for the response function cannot always be logistic. That limitation does not hinder calibration weighting when performed in two steps: from the respondent sample to the full sample to remove the response bias and then from the full sample to the population to decrease variance. There are potential efficiency advantages from using the two-step approach as well even when the calibration variables employed in each step is a subset of the calibration variables in the single step. Simultaneous mean-squared-error estimation using linearization is possible, but more complicated than when calibrating in a single step.

Key Words: Probability sampling; Response model; Prediction model; Double protection; Simultaneous variance estimation.

1 Introduction

Survey sampling is a tool used primarily for estimating the parameters of a finite population based on a randomly drawn sample of its members. Probability samples come with design (sampling) weights, which are often the inverses of the individual member selection probabilities. As long as each population element has a positive selection probability, it is a simple matter to produce an estimator for the population total of a survey variable that is unbiased with respect to the probability-sampling mechanism. The ratio of two unbiased estimators of totals or any other smooth function of estimated totals, while not necessarily unbiased, is asymptotically unbiased and often consistent since its relative variance, like its relative bias, tends to zero as the sample size grows arbitrarily large.

Deville and Särndal (1992) introduced calibration weighting as a tool for adjusting design weights in such a way that the weighted sums of certain "calibration" variables equal their known (or better-estimated) population totals. As a consequence of these *calibration equations* holding, the standard error of an estimated total for a variable without a known population total is often reduced while remaining nearly (i.e., asymptotically) unbiased under the probability sampling mechanism.

Although originally developed to reduce standard errors, calibration weighting has also been used to remove selection biases resulting from unit nonresponse under certain assumptions (e.g., Folsom 1991; Fuller, Loughin and Baker 1994; Lundström and Särndal 1999; Folsom and Singh 2000). To this end, whether (or not) an element selected for the sample responds to a survey is treated as an additional phase of Poisson random sampling with unknown, but positive, selection probabilities. Calibration weighting estimates these Poisson selection probabilities implicitly and produces estimated totals that are nearly

1. Phillip S. Kott, Senior Research Statistician, RTI International, Rockville, Maryland 20852, U.S.A. E-mail: pkott@rti.org; Dan Liao, Research Statistician, RTI International, Rockville, Maryland 20852, U.S.A.

unbiased under the combined sample- and response-selection mechanisms, which is often called the "quasi-sampling design". See Oh and Scheuren (1983).

An important *caveat* is that although the sample-selection mechanism is fully under the control of the statistician, the response-selection mechanism is unknown. The response mechanism is assumed to have a particular form, and the failure of this assumption can result in biased estimators.

An alternative justification for calibration weighting involves a different type of modeling. It is easy to show that calibration weighting produces an estimator that is unbiased under a linear prediction (outcome) model if the expected value of the survey variable under the prediction model is a linear function of the calibration variables so long as the sampling and response mechanisms are ignorable, that is to say, the same prediction model applies whether or not the population element is sampled or whether it responds when sampled.

Unlike the selection model governing the response mechanism, it is possible for the linear prediction model to hold for one survey variable and not another. That is why most survey samplers prefer to assume a *selection model* when adjusting for unit nonresponse. Nevertheless, it is reassuring to know that if *either* model is correct, then the estimated total is nearly unbiased (i.e., has a relative bias that vanishes asymptotically), a property Kim and Park (2006) called "double protection" against nonresponse bias.

It is possible to simultaneously remove the selection bias and decrease standard error under the probability-sample mechanism in a single step by adjusting the design weights of unit respondents so that the estimated totals for a set of calibration variables equal their known population totals. Nevertheless, there are reasons for preferring the use of two calibration-weighting steps even when the sets of calibration variables used in both steps are the same or a subset of the calibration variables in a single step: the first step from the respondent sample to the original sample to remove selection bias and the second from the original sample to the population to decrease the variances of the resulting estimators.

Although Folsom and Singh (2000) and others have pointed out that calibration weighting can also be used to remove the selection bias due to under- or over-coverage of the sampling frame, we will direct our attention here on a single-stage sample drawn from a complete list frame without duplication. That is to say, we will assume that the sampling frame is identical to the target population (i.e., each population unit is listed on the frame).

The paper is structured as follows. Section 2 reviews some background theory on calibration weighting. Section 3 introduces a slightly new variance estimator that, like the variance estimator in Kott (2006), can be used to measure both the mean squared error of a calibration-weighted estimator under the quasi-sampling design and the variance under either the prediction model or the combination of the prediction model and original sampling mechanism, thus making the double protection against nonresponse bias arguably more useful for inference. The variance estimator in Kott applies only when calibrating to the population. Here we follow Folsom and Singh (2000) and allow the possibility that calibration is to the original sample.

Section 4 discusses the limitations of calibrating weighting in a single step and develops some theory for a two-step approach. Although our main purpose here is to argue the benefits of using two steps even when similar sets of calibration variables are employed in both steps, the calibration estimator we treat in this section is broader. Section 5 describes the results of some simulation experiments, while Section 6 offers a few concluding remarks.

2 One-step calibration weighting

2.1 Calibration weighting and unit nonresponse

In the absence of nonresponse (or frame errors), calibration weighting is a sampling-weight-adjustment method that creates a set of weights $\{w_k; k \in S\}$, asymptotically close to the original design weights, $d_k = 1/\pi_k$, that satisfy a set of *calibration equations* (one for each component of $\mathbf{z}_k$) :

$$\sum_S w_k \mathbf{z}_k = \sum_U \mathbf{z}_k,$$

where S denotes the sample, π_k the sample-selection probability of unit k, U the population of size $N, \mathbf{z}_k$ a vector with P components each having a known population total, and $\sum_A$ means $\sum_{k \in A}$.

Kott (2009) describes a conservative set of mild conditions under which $t_y = \sum_S w_k y_k$ is a nearly unbiased estimator for the population total $T_y = \sum_U y_k$ (i.e., the relative bias of t_y is asymptotically zero). Most importantly, each $\pi_k N/n$ is assumed to be bounded from below by a positive value as N and the (expected) sample size, n, grow arbitrarily large (we add the parenthetical "expected" in case the sample size is random).

In addition, the first four central population moments of each component of $\mathbf{z}_k$ is assumed to be bounded from above, while $N^{-1} \sum_U \mathbf{z}_k \mathbf{z}_k^T$ converges to a positive definite matrix.

Using calibration-weighting will tend to reduce mean squared error relative to the expansion estimator, $t_y^E = \sum_S d_k y_k$, when y_k is correlated with some components of $\mathbf{z}_k$. One should keep in mind, however, that most surveys have many y_k's.

A simple way to compute calibration weights is linearly with the following formula:

$$\begin{aligned} w_k &= d_k \left[1 + \left(\sum_U \mathbf{z}_j - \sum_S d_j \mathbf{z}_j\right)^T \left(\sum_S d_j \mathbf{z}_j \mathbf{z}_j^T\right)^{-1} \mathbf{z}_k\right] \\ &= d_k \left[1 + \mathbf{g}^T \mathbf{z}_k\right]. \end{aligned}$$

Fuller et al. (1994) and later Lundström and Särndal (1999) argued that this linear calibration can also be used to handle unit nonresponse. The sample S is replaced by the respondent sample R, while

$$\mathbf{g} = \left[(1-\theta)\left(\sum_U \mathbf{z}_j - \sum_R d_j \mathbf{z}_j\right)^T + \theta\left(\sum_S d_j \mathbf{z}_j - \sum_R d_j \mathbf{z}_j\right)^T\right]\left(\sum_R d_j \mathbf{z}_j \mathbf{z}_j^T\right)^{-1},$$

depending on whether the respondent sample is *calibrated to the population* $(\theta = 0)$ or *calibrated to the original sample* $(\theta = 1)$. Either way, the estimate is nearly unbiased under the quasi-sample-design that treats response as a second phase of random sampling so long as each unit's probability of response has the form:

$$p_k = 1/\left(1 + \boldsymbol{\gamma}^T \mathbf{z}_k\right), \tag{2.1}$$

and $\mathbf{g}$ is a consistent estimator for the unknown parameter vector $\boldsymbol{\gamma}$ in equation (2.1).

The problem with the response function in equation (2.1) is that the implicit estimator for p_k, $\hat{p}_k = 1/(1 + \mathbf{g}^T\mathbf{z}_k)$ can be negative. A nonlinear form of calibration weighting avoiding this possibility was suggested by Kott and Liao (2012) based on the generalized exponential form of Folsom and Singh (2000). It uses Newton's method (iterative Taylor-series approximations) to find a $\mathbf{g}$ such that the calibration equation (from here on, we refer to the vector of component calibration equations as the calibration equation):

$$\sum_R w_k\mathbf{z}_k = \sum_R d_k\alpha(\mathbf{g}^T\mathbf{z}_k)\mathbf{z}_k = (1-\theta)\sum_U \mathbf{z}_k + \theta\sum_S d_k\mathbf{z}_k \tag{2.2}$$

holds, where $\theta = 0$ or 1,

$$\alpha(\mathbf{g}^T\mathbf{z}_k) = \frac{\ell + \exp(\mathbf{g}^T\mathbf{z}_k)}{1 + \exp(\mathbf{g}^T\mathbf{z}_k)/u}, \tag{2.3}$$

ℓ, the lower bound of $\alpha(\cdot)$, is nonnegative (so that calibration weights are likewise nonnegative), and the upper bound of $\alpha(\cdot)$, $u > \ell$, can be either finite or infinite.

Although there are other reasonable forms the *weight-adjustment function* $\alpha(\mathbf{g}^T\mathbf{z}_k)$ can take, we will restrict our attention to functions in the form in equation (2.3). This is a generalization of both raking where $\ell = 0, u = \infty$, and the implicit estimation of a logistic response model, where $\ell = 1, u = \infty$. In Deming and Stephan's original (1940) iterative-proportional-fitting algorithm for raking, the components of $\mathbf{z}_k$ were restricted to indicator functions. We use "raking" more broadly here to mean calibration weighting with a weight-adjustment function of the form $\alpha(\mathbf{g}^T\mathbf{z}_k) = \exp(\mathbf{g}^T\mathbf{z}_k)$.

When $\ell < 1$, equation (2.3) becomes the generalized-raking adjustment introduced in Deville and Särndal (1992) and discussed further in Deville, Särndal and Sautory (1993). Generalized raking not only lets the components of $\mathbf{z}_k$ be continuous but also allows the range of the $\alpha(\mathbf{g}^T\mathbf{z}_k)$ to be constrained between a positive ℓ and a (possibly) finite u.

Deville and Särndal (1992) required $\alpha(0) = \alpha'(0) = 1$. Since the authors were not treating samples with nonresponse (or incorrect frames), $\mathbf{g}^T\mathbf{z}_k$ needed to converge to 0 and $\alpha(\mathbf{g}^T\mathbf{z}_k)$ to 1 as the (expected) sample size grew arbitrarily large. When adjusting design weights for nonresponse, however, setting $\ell \geq 1$ is a more sensible strategy, so that the implicit estimated probability of response does not exceed 1.

Although the original definition of calibration weighting in Deville and Särndal (1992) involved minimizing the differences between the w_k and d_k in R as measured by some loss function, later formulations (e.g., Estevao and Särndal 2000) removed the loss function from the definition. Forcing w_k and d_k to be close makes little sense when calibration weighting is used to adjust for unit nonresponse since if a sampled k has a relatively small probability of response, then the difference between w_k and d_k *should* be relatively large.

Rather than assuming a response model with a particular functional form, an alternative justification for using calibration weighting as a mean of removing unit-nonresponse bias assumes a prediction model in which the survey variable y_k is itself a random variable such that $\mathrm{E}(y_k|\mathbf{z}_k) = \mathbf{z}_k^T\boldsymbol{\beta}$ for some unknown $\boldsymbol{\beta}$ whether or not k is sampled or whether it responds when sampled. Kott (2006) and others have

observed the calibration-weighted estimator for $T_y = \sum_U y_k$ will be nearly unbiased under the prediction model when calibration is done to the population (when $\theta = 0$ in equation (2.2)) and under the combination of the prediction model and the original sample-selection mechanism when calibration is done to the original sample (when $\theta = 1$).

The property that a calibration-weighted estimator is nearly unbiased in some sense when *either* an assumed response model *or* an assumed prediction model holds has been called "double protection against nonresponse bias" by Kim and Park (2006). It is known as "double robustness" in the biostatics literature (Bang and Robins 2005) and attributed to Robins, Rotnitzky and Zhao (1994), which dealt with item rather than unit nonresponse.

The distribution of $y_k|\mathbf{z}_k$ under the prediction model is often assumed to be the same for sampled and nonsampled population members. That is to say, the sampling mechanism is assumed to be *ignorable*. In addition, the distribution of $y_k|\mathbf{z}_k$ is often assumed to be the same whether or not a population member responds when sampled, that is, that the response mechanism is also assumed to be ignorable (Little and Rubin 2002). Here, we make weaker analogous assumptions under the prediction model, namely, that $\mathrm{E}(y_k|\mathbf{z}_k)$ does not depend on whether k is sampled or when sampled responds. Let us say that the sampling and response mechanisms are assumed to be "first-moment ignorable".

2.2 Instrumental variables

Deville (2000) observed that instrumental-variable calibration can be used to adjust for potential nonresponse bias by assuming a response model that depended on $\mathbf{x}_k$,

$$p_k = \left[\alpha\left(\boldsymbol{\gamma}^T\mathbf{x}_k\right)\right]^{-1} = \frac{1+\exp\left(\boldsymbol{\gamma}^T\mathbf{x}_k\right)/u}{\ell+\exp\left(\boldsymbol{\gamma}^T\mathbf{x}_k\right)}, \tag{2.4}$$

but fitting calibration equations with $\mathbf{z}_k$:

$$\sum_R w_k\mathbf{z}_k = \sum_R d_k\alpha\left(\mathbf{g}^T\mathbf{x}_k\right)\mathbf{z}_k = (1-\theta)\sum_U \mathbf{z}_k + \theta\sum_S d_k\mathbf{z}_k, \tag{2.5}$$

where the $\mathbf{g}$ satisfying equation (2.5) with $\theta = 0$ or 1 a consistent estimator of unknown parameter vector $\boldsymbol{\gamma}$ in equation (2.4). Some mild conditions are needed for this. Sufficient are the following: $N^{-1}\sum_R d_k\alpha\left(\boldsymbol{\gamma}^T\mathbf{x}_k\right)\mathbf{z}_k$ is a consistent and bounded estimator for $N^{-1}[(1-\theta)\sum_U \mathbf{z}_k + \theta\sum_S d_k\mathbf{z}_k]$, $\alpha(\phi)$ is everywhere twice differentiable, and $N^{-1}\sum_R d_k\alpha'(\phi)\mathbf{z}_k\mathbf{x}_k^T$ is always invertible and bounded as the sample grows arbitrarily large.

Let $R_k = 1$ when $k \in R, 0$ otherwise. It is not hard to show that

$$\begin{aligned}\mathbf{g}-\boldsymbol{\gamma} = &-\left(\sum_S d_kR_k\alpha'(c_k)\mathbf{z}_k\mathbf{x}_k^T\right)^{-1}\left\{\sum_S d_kR_k\alpha\left(\boldsymbol{\gamma}^T\mathbf{x}_k\right)\mathbf{z}_k - \left[(1-\theta)\sum_U \mathbf{z}_k + \theta\sum_S d_k\mathbf{z}_k\right]\right\}\\ &-\left(N^{-1}\sum_S d_kR_k\alpha'(c_k)\mathbf{z}_k\mathbf{x}_k^T\right)^{-1}\left\{N^{-1}\sum_S d_kR_k\alpha\left(\boldsymbol{\gamma}^T\mathbf{x}_k\right)\mathbf{z}_k - N^{-1}\left[(1-\theta)\sum_U \mathbf{z}_k + \theta\sum_S d_k\mathbf{z}_k\right]\right\}\end{aligned}$$

for some c_k between $\mathbf{g}^T\mathbf{x}_k$ and $\boldsymbol{\gamma}^T\mathbf{x}_k$, as Kott and Liao (2012) demonstrated when $\mathbf{x}_k = \mathbf{z}_k$.

Deville also noted that it is possible for components of the $\mathbf{x}_k$ to be survey variables with values known only for respondents. Chang and Kott (2008) extended the notion of calibration weighting to allow the dimension of the $\mathbf{z}_k$-vector to be greater than that of the $\mathbf{x}_k$-vector. We will *not* treat either possibility in the following sections.

Kim and Shao (2013) in treating nonignorable nonresponse call the components of $\mathbf{z}_k$ not wholly functions of the components of $\mathbf{x}_k$ "instrumental variables". To limit future confusion, we will henceforth use to term "model variables" to refer to the components of $\mathbf{x}_k$.

3 Variance estimation for the one-step calibration estimator

In this section, we let

$$t_y = \sum_R w_k y_k = \sum_R d_k \alpha\left(\mathbf{g}^T \mathbf{x}_k\right) y_k$$

be the calibration-weighted estimator for T_y, where $w_k = d_k \alpha\left(\mathbf{g}^T \mathbf{x}_k\right)$ when $k \in R$ is the calibration weight, and w_k is conveniently defined to be 0 when $k \notin R$. The weight-adjustment function $\alpha(\cdot)$ is defined implicitly by equation (2.4), and $\mathbf{g}$ is again chosen so that the calibration equation (2.5) holds for either $\theta = 0$ or 1.

We propose the following estimator for the variance t_y :

$$v\left(t_y\right) = \sum_{k,j \in S} \left(1 - \frac{\pi_k \pi_j}{\pi_{kj}}\right)\left[d_k\left(\theta \mathbf{z}_k^T \mathbf{b} + \alpha_k e_k\right)\right]\left[d_j\left(\theta \mathbf{z}_j^T \mathbf{b} + \alpha_j e_j\right)\right] + \sum_{k \in R} d_k\left(\alpha_k^2 - \alpha_k\right) e_k^2, \quad (3.1)$$

where π_{kj} is the joint selection probability of k and j under the original sampling design, $\pi_{kk} = \pi_k = 1/d_k$, $\pi_k = \alpha\left(\mathbf{g}^T \mathbf{x}_k\right)$ when $k \in R$ and 0 otherwise,

$$\mathbf{b} = \left[\sum_R d_k \alpha'\left(\mathbf{g}^T \mathbf{x}_k\right) \mathbf{x}_k \mathbf{z}_k^T\right]^{-1} \sum_R d_k \alpha'\left(\mathbf{g}^T \mathbf{x}_k\right) \mathbf{x}_k y_k, \quad (3.2)$$

and $e_k = y_k - \mathbf{z}_k^T \mathbf{b}$. We will show that $v\left(t_y\right)$ in equation (3.1) can be nearly unbiased in some sense if *either* a response model (Section 3.1) *or* prediction model holds (Section 3.2).

The variance estimator in equation (5.2) of Kott (2006) is identical to $v\left(t_y\right)$ in equation (3.1) when $\theta = 0$. The variance estimator in Kim and Haziza (2014) is also similar. Their prediction model is more general than the linear prediction model considered here.

This variance estimator $v\left(t_y\right)$ presupposes that the original sampling design is such that each element can only be drawn once. In Section 3.1, we see that when the probabilities of response are independent (Poisson), then under mild assumptions, $v\left(t_y\right)$ is a nearly unbiased estimator of the mean squared error of t_y under the quasi-sampling design whether or not the prediction model, $\mathrm{E}\left(y_k \left| \mathbf{x}_k, \mathbf{z}_k \right.\right) = \mathbf{z}_k^T \boldsymbol{\beta}$, holds.

In Section 3.2, $v(t_y)$ is shown to be a nearly unbiased estimator for the combined prediction-model and original-sampling-design variance of t_y as an estimator for T_y whether or not the response model in equation (2.4) holds. Thus, $v(t_y)$ can be called a "simultaneous variance estimator".

3.1 Variance estimation under the response model

For ease of exposition we will assume that the response model in equation (2.4) with a finite u holds. Sufficient conditions for $v(t_y)$ to be a nearly unbiased estimator for the mean squared error of t_y (by which the bias converges to 0 as the sample size grows arbitrary large) are

$$\pi_{kj} \geq B_0 > 0 \tag{3.3}$$

$$\sum_{j=1}^{N} \left| \frac{\pi_{kj}}{\pi_k \pi_j} - 1 \right| \leq B_1 < \infty \text{ for every } k, \tag{3.4}$$

$$\frac{\sum_{j=1}^{N} \psi_j^r}{N} \leq B_2 < \infty \text{ where } \psi_j \text{ is } y_j \text{ or any component of } \mathbf{x}_j \text{ or } \mathbf{z}_j, \text{while } r = 1 \text{ or } 2, \tag{3.5}$$

and $N^{-1} \sum_R d_k \alpha'(\mathbf{g}^T \mathbf{x}_k) \mathbf{z}_k \mathbf{x}_k^T$ is of full rank and is bounded in probability as the sample size grows arbitrarily large.

From these, $\alpha'(\phi) = (1 - \alpha(\phi)/u)\exp(\phi)/[(1 + \exp(\phi)/u)]$ being bounded when u is finite, and the Cauchy-Schwarz inequality $\left((\sum a_k b_k)^2 \leq \sum a_k^2 \sum b_k^2\right)$, it is not hard to see not only that $\mathbf{g}$ is a consistent estimator for $\boldsymbol{\gamma}$, but also that $\mathbf{b}$ in equation (3.2) (which can be rendered $\mathbf{b} = \left[N^{-1} \sum_R d_k \alpha'(\mathbf{g}^T \mathbf{x}_k) \mathbf{x}_k \mathbf{z}_k^T\right]^{-1} N^{-1} \sum_R d_k \alpha'(\mathbf{g}^T \mathbf{x}_k) \mathbf{x}_k y_k$) has a probability limit, call it $\mathbf{b}^*$, whether or not the prediction model holds. Moreover, both $\mathbf{b} - \mathbf{b}^*$ and $\mathbf{g} - \boldsymbol{\gamma}$ are $\mathbf{O}_p(1/\sqrt{n})$.

Observe that

$$\begin{aligned}
(t_y - T_y)/N &= \theta\left(\sum_S d_k \mathbf{z}_k^T \mathbf{b}^* - \sum_U \mathbf{z}_k^T \mathbf{b}^*\right)/N \\
&+ \left[\sum_R d_k \alpha(\mathbf{g}^T \mathbf{x}_k) e_k^* - \sum_R d_k \alpha(\boldsymbol{\gamma}^T \mathbf{x}_k) e_k^*\right]/N \\
&+ \left[\sum_R d_k \alpha(\boldsymbol{\gamma}^T \mathbf{x}_k) e_k^* - \sum_U e_k^*\right]/N,
\end{aligned}$$

where $e_k^* = y_k - \mathbf{z}_k^T \mathbf{b}^*$. The insertion of the $\alpha'(\cdot)$ into the "regression coefficient" $\mathbf{b}$ allows us to ignore the contribution to quasi-design mean squared error of the second term in this sum, $Q = \sum_R d_k \left[\alpha(\mathbf{g}^T \mathbf{x}_k) - \alpha(\boldsymbol{\gamma}^T \mathbf{x}_k)\right] e_k^*/N$. That is because $\sum_R d_k \alpha'(\boldsymbol{\gamma}^T \mathbf{x}_k) \mathbf{x}_k e_k = 0$ is true by definition, which implies $\sum_R d_k \alpha'(\boldsymbol{\gamma}^T \mathbf{x}_k) \mathbf{x}_k e_k^*$ is $\mathrm{O}_p(1/\sqrt{n})$ under our assumptions. Moreover, since $\alpha(\mathbf{g}^T \mathbf{x}_k) - \alpha(\boldsymbol{\gamma}^T \mathbf{x}_k) = \alpha'(c_k)(\mathbf{g} - \boldsymbol{\gamma})^T \mathbf{x}_k$ is also $\mathrm{O}_p(1/\sqrt{n})$, $Q = (\mathbf{g} - \boldsymbol{\gamma})^T \sum_R d_k \alpha'(c_k) \mathbf{x}_k e_k^*$ is $\mathrm{O}_p(1/n)$, which is asymptotically ignorable relative to the two $\mathrm{O}_p(1/\sqrt{n})$ components of $(t_y - T_y)/N$.

With the contribution of Q eliminated from consideration, an idealized, but not calculable, nearly unbiased estimator for the quasi-design mean squared error of t_y is

$$v_{I1}(t_y) = \sum_{k,j\in S}\left(1-\frac{\pi_k\pi_j}{\pi_{kj}}\right)\left[d_k\left(\theta\mathbf{z}_k^T\mathbf{b}^* + e_k^*\right)\right]\left[d_j\left(\theta\mathbf{z}_j^T\mathbf{b}^* + e_j^*\right)\right] + \sum_{k\in R}\left(\frac{d_k e_k^*}{p_k}\right)^2(1-p_k), \quad (3.6)$$

where the first term on the right estimates the mean squared error before nonresponse (if any) and the second the added variance due to nonresponse.

An alternative nearly unbiased idealized mean squared error estimator, closer to being calculable, is

$$v_{I2}(t_y) = \sum_{k,j\in S}\left(1-\frac{\pi_k\pi_j}{\pi_{kj}}\right)\left[d_k\left(\theta\mathbf{z}_k^T\mathbf{b}^* + \frac{R_k}{p_k}e_k^*\right)\right]\left[d_j\left(\theta\mathbf{z}_j^T\mathbf{b}^* + \frac{R_j}{p_j}e_j^*\right)\right] + \sum_{k\in R} d_k\left(\frac{e_k^*}{p_k}\right)^2(1-p_k), \quad (3.7)$$

where again $R_k = 1$ when $k \in R, 0$ otherwise. Since the $(R_k/p_k)e_k^*$ are independent under the response model with mean e_k^* and variance $(e_k^*/p_k)^2 p_k(1-p_k)$, $\mathrm{E}\left[(R_k/p_k)e_k^*(R_j/p_j)e_j^*\right] = e_k^* e_j^*$ when $k \neq j$. By contrast, the following holds when $k = j$:

$$\begin{aligned}(1-\pi_k)E\left[\left(d_k\frac{R_k}{p_k}e_k^*\right)^2\right] &= (1-\pi_k)\left[(d_k e_k^*)^2 + \left(\frac{d_k e_k^*}{p_k}\right)^2 p_k(1-p_k)\right] \\ &= (1-\pi_k)(d_k e_k^*)^2 + \left(\frac{d_k e_k^*}{p_k}\right)^2 p_k(1-p_k) - d_k\left(\frac{e_k^*}{p_k}\right)^2 p_k(1-p_k).\end{aligned}$$

The first summation on the right-hand side of equation (3.7) has terms where $k \neq j$ and terms where $k = j$, the latter of which causes the second summation in (3.7) to differ from the second summation on the right-hand side of equation (3.6). Note that the expectation under the response model of $\sum_R d_k(e_k^*/p_k)^2(1-p_k)$ in the second summation on the right-hand side of (3.7) is $\sum_S d_k(e_k^*/p_k)^2 p_k(1-p_k)$.

Finally, $v_{I2}(t_y)$ can be replaced by the asymptotically identical, but computable, $v(t_y)$ in equation (3.1) since $\sum_{j\in S}(1-\pi_k\pi_j/\pi_{kj})$ is bounded for all k under assumptions (3.3) and (3.4), allowing e_k and α_k to be substituted for the unknown e_k^* and $1/p_k$, respectively (because $e_k^* - e_k$ and $\alpha_k - 1/p_k$ are $\mathrm{O}_p(1/\sqrt{n})$ for all k).

3.2 Variance estimation under the prediction model

Matters are a bit simpler when we assume a prediction model holds but not necessarily the response model in equation (2.4). Suppose $\mathrm{E}(y_k|\mathbf{x}_k, \mathbf{z}_k) = \mathbf{z}_k^T\boldsymbol{\beta}$, whether or not k is sampled or responds when sampled, and the $\varepsilon_k = y_k - \mathbf{z}_k^T\boldsymbol{\beta}$ are uncorrelated random variables with variances equal to $\sigma_k^2 = \mathbf{z}_k^T\boldsymbol{\eta}$, where $\boldsymbol{\eta}$ need not be specified other than having finite components.

The mean squared error of t_y as an estimator for T_y under that prediction model is the sum of the prediction variance of t_y as an estimator for T_y, $\sum_R \left(w_k^2 - w_k\right)\sigma_k^2$ (see, for example, Kott 2009, page 69), and the squared bias, $\left(\sum_S \mathbf{x}_k^T\boldsymbol{\beta} - \sum_U \mathbf{x}_k^T\boldsymbol{\beta}\right)^2$, the latter being zero when $\theta = 0$. The combined variance of t_y as an estimator for T_y under the prediction model and original sample design is

$$V_C = \theta \mathrm{Var}_D\left(\sum_S \mathbf{x}_k^T\boldsymbol{\beta}\right) + \mathrm{E}_D\left[\sum_S \left(w_k^2 - w_k\right)\sigma_k^2\right],$$

where the subscript D denotes that the operation (variance or expectation) is with respect to the original sampling design. Recall $w_k = 0$ for $k \neq R$.

To see that $v\left(t_y\right)$ in equation (3.1) provides a nearly unbiased estimator for V_C, observe first that

$$e_k = y_k - \mathbf{z}_k^T\mathbf{b} = \varepsilon_k - \mathbf{z}_k^T\left[N^{-1}\sum_R d_j\alpha'\left(\mathbf{g}^T\mathbf{x}_j\right)\mathbf{x}_j\mathbf{z}_j^T\right]^{-1} N^{-1}\sum_R d_j\alpha'\left(\mathbf{g}^T\mathbf{x}_j\right)\mathbf{x}_j\varepsilon_j.$$

Let $\delta_{kj} = 1$ when $k = j$ and 0 otherwise. Because the ε_k are uncorrelated, and $\mathrm{E}\left(\varepsilon_k^2\right) = \sigma_k = \mathbf{z}_k^T\boldsymbol{\eta}$, it is now not hard to show that $\mathrm{E}\left(e_k e_j\right) = \delta_{kj}\sigma_k^2 + \mathrm{O}(1/n)$ for almost every k, j pair under the prediction model when $N^{-1}\sum_R d_k\alpha'\left(\mathbf{g}^T\mathbf{x}_k\right)\mathbf{z}_k\mathbf{x}_k^T$ converges to an invertible matrix, and assumptions (3.3), (3.4), and

$$\frac{\sum_{j=1}^{N}\psi_j^r}{N} \leq B_2 < \infty \text{ where } \psi_j \text{ is any component of } \mathbf{x}_j \text{ or } \mathbf{z}_j, \text{ and } r = 1, 2, 3, \text{ or } 4, \qquad (3.8)$$

hold. Observe that the change from the assumptions in (3.5) to (3.8) makes the relative bias of $v\left(t_y\right)$ as an estimator for V_C (or $\sum_R \left(w_k^2 - w_k\right)\sigma_k^2$ when $\theta = 0$) $\mathrm{O}(1/n)$ rather than $\mathrm{O}\left(1/\sqrt{n}\right)$.

4 Two-step calibration weighting

4.1 Calibration weighting in two steps

In practice, the components of $\mathbf{x}_k$ are often 0/1 group-membership identifiers, and the groups are mutually exclusive and exhaustive. In that situation, $\mathbf{g}^T\mathbf{x}_k$ can only take on P values. *Almost* any weight-adjustment function, $\alpha\left(\mathbf{g}^T\mathbf{x}_k\right)$, will yield equivalent results. An example is the linear function, $\alpha\left(\mathbf{g}^T\mathbf{x}_k\right) = 1 + \mathbf{g}^T\mathbf{x}_k$, of Lundström and Särndal (1999).

One popular weight-adjustment function that sometimes *cannot* be used (note the italicized "almost" in the previous paragraph) is $\alpha\left(\mathbf{g}^T\mathbf{x}_k\right) = 1 + \exp\left(\mathbf{g}^T\mathbf{x}_k\right)$, which assumes response is a logistic function of $\mathbf{x}_k$. The problem is that this weight-adjustment function cannot return values less than unity. We noted in the previous section, that sometimes one may need α_k to be less than 1. A routine that tries to use $\alpha\left(\mathbf{g}^T\mathbf{x}_k\right) = 1 + \exp\left(\mathbf{g}^T\mathbf{x}_k\right)$ and fit the calibration equations will fail.

This can be a particular problem when assuming a logistic response model and trying to calibrate to the population in a single step. There may be a component of $\mathbf{z}_k$, say z_{ka}, that is always nonnegative, but the original sample and response set are such that $\sum_R d_k z_{ka} > \sum_U z_{ka}$ even though $\sum_R d_k z_{ka}$ cannot exceed $\sum_S d_k z_{ka}$. Thus, calibrating to the population will always fail because no α_k can be less than 1.

Calibrating to the original sample, by contrast, need not fail, since $\sum_R d_k z_{ka} \leq \sum_S d_k z_{ka}$. This suggest that one calibrates first to the original sample, which removes the response bias if the assumed response model holds, and then to the population, which removes the response bias if the prediction model holds. Estevao and Särndal (2002) discuss a variety of ways to calibrate in steps, but we focus on a single method here.

A second advantage of calibration weighting in two steps can be realized even when the calibration variables used in both steps are the same or a subset of those used in the single step. This happens when the response model holds, and the linear prediction model is only roughly true. Some version or "optimal" estimation can then be used in the second calibration-weighting step to increase efficiency. Rao (1994) introduced the notion of the optimal regression estimator. It was put into calibration-weighting form and discussed further in Bankier (2002) and Kott (2009, Section 4.2). Detail and how this can be done are provided in Sections 4.2 and 5.

4.2 Estimation and variance estimation when calibrating in two steps

In this subsection, we start with a fairly general two-step calibration estimator for a total and then address estimating its variance. The first calibration-weighting step, which is to the original sample, employs $\mathbf{x}_{1k}$ as the vector of response-model variables and $\mathbf{z}_{1k}$ as the calibration vector. Each has P_1 components. The weight-adjustment function has the form described in equation (2.4) with $\mathbf{g}_1$ now replacing $\mathbf{g}$. The calibration equation is $\sum_R d_k \alpha\left(\mathbf{g}_1^T \mathbf{x}_{1k}\right) \mathbf{z}_{1k} = \sum_S d_k \mathbf{z}_{1k}$.

The second calibration-weighting step, which is to the population, employs $\mathbf{x}_{2k}$ and $\mathbf{z}_{2k}$, each with P_2 components. The nonresponse bias under the response model is removed in the first step. For the weight-adjustment function for the second step, we propose using

$$h_k\left(\mathbf{g}_2^T \mathbf{x}_{2k}\right) = \frac{\ell_k + \exp\left(\mathbf{g}_2^T \mathbf{x}_{2k}\right)}{1 + \exp(\mathbf{g}_2^T \mathbf{x}_{2k})/u_k}, \tag{4.1}$$

where $u_k > \ell_k > 0$ may be set almost at whim (but see below). The right-hand side of equation (4.1) can vary across the k (and so can depend on d_k and α_k), yet $h_k(0) = h_k'(0) = 1$, making it asymptotically indistinguishable from the linear function: $1 + \mathbf{g}_2^T \mathbf{x}_{2k}$. For simplicity, we will call $h_k\left(\mathbf{g}_2^T \mathbf{x}_{2k}\right)$ and $h_k'\left(\mathbf{g}_2^T \mathbf{x}_{2k}\right), h_k$ and h_k' respectively. From a quasi-sampling-design viewpoint, both are asymptotically identical to unity. The second calibration equation is $\sum_S d_k h_k\left(\mathbf{g}_2^T \mathbf{x}_{2k}\right) \mathbf{z}_{2k} = \sum_U \mathbf{z}_{2k}$. Because this equation must hold, there are limits on the available choices for u_k and ℓ_k in equation (4.1).

A good simultaneous variance estimator for $t_y = \sum_R w_k y_k = \sum_R d_k \alpha\left(\mathbf{g}_1^T \mathbf{x}_{1k}\right) h_k\left(\mathbf{g}_2^T \mathbf{x}_{2k}\right) y_k$ is (as we shall see)

$$v\left(t_y\right) = \sum_{k,j\in S}\left(1-\frac{\pi_k\pi_j}{\pi_{kj}}\right)\left[d_k\left(\mathbf{z}_{1k}^T\mathbf{b}_1+\alpha_k h_k e_{1k}\right)\right]\left[d_j\left(\mathbf{z}_{1j}^T\mathbf{b}_1+\alpha_j h_j e_{1j}\right)\right] + \sum_{k\in R} d_k\left(h_k^2\alpha_k^2-h_k\alpha_k\right)e_{1k}^2, \tag{4.2}$$

where

$$e_{2k} = y_k - \mathbf{z}_{2k}^T\left(\sum_S d_j\alpha_j h_j'\mathbf{x}_{2j}\mathbf{z}_{2j}^T\right)^{-1}\sum_S d_j\alpha_j h_j'\mathbf{x}_{2j}y_j, \tag{4.3}$$

$$\mathbf{b}_1 = \left(\sum_S d_f\alpha_f'\mathbf{x}_{1f}\mathbf{z}_{1f}^T\right)^{-1}\sum_S d_f\alpha_f' h_f\mathbf{x}_{1f}e_{2f}, \tag{4.4}$$

and

$$e_{1k} = e_{2k} - \mathbf{x}_{1k}^T\mathbf{b}_1. \tag{4.5}$$

Let $\mathbf{x}_k$ now be the vector composed of the non-duplicated components of $\mathbf{x}_{1k}$ and $\mathbf{x}_{2k}$ and define $\mathbf{z}_k$ analogously. Sufficient conditions for (4.2) to be a simultaneous variance estimator include the corresponding components of equation (4.1) depending on whether either the response model in equation (2.4) holds with $\mathbf{x}_{1k}$ replacing $\mathbf{x}_k$ or the prediction model is $\mathrm{E}\left(y_k|\mathbf{x}_k,\mathbf{z}_k\right) = \mathbf{z}_{2k}^T\boldsymbol{\beta}_2$, whether or not k is sampled or responds if sampled, and the $\varepsilon_{2k} = y_k - \mathbf{z}_{2k}^T\boldsymbol{\beta}_2$ are uncorrelated random variables with variances equal to $\sigma_{2k}^2 = \mathbf{z}_{2k}^T\boldsymbol{\eta}_2$, where $\boldsymbol{\eta}_2$ need not be specified other than having finite components. Now, both $N^{-1}\sum_R d_k\alpha'\left(\mathbf{g}_1^T\mathbf{x}_{1k}\right)\mathbf{z}_{1k}\mathbf{x}_{1k}^T$ and $N^{-1}\sum_R d_k h_k'\left(\mathbf{g}_2^T\mathbf{x}_{2k}\right)\mathbf{z}_{2k}\mathbf{x}_{2k}^T$ are assumed to be of full rank and bounded as the sample size grows arbitrarily large.

The variance estimator in equation (4.2) is almost the same as the estimator in (3.1): $\mathbf{x}_k$ has been replaced with $\mathbf{x}_{1k}$ and $\mathbf{z}_k$ with $\mathbf{z}_{1k}$, while $h_k e_{2k}$ substitutes for y_k (we will get to a small difference shortly). Observe that e_{2k} is effectively an expression of the "residual" from the second calibration-weighting step. This residual is multiplied by the weight-adjustment factor h_k, which is asymptotically unity from the quasi-sampling-design-based perspective and a constant from the prediction-model viewpoint. The product is then used to create the first-step "regression-coefficient" $\mathbf{b}_1$ in equation (4.4) and its accompanying "residual" e_{1k} in equation (4.5). We do the second step regression first because $t_y - T_y = \sum_R w_k y_k - \sum_U y_k = \sum_R w_k e_{2k} - \sum_U e_{2k}$.

It is for estimating the prediction model of t_y as an estimator of $T_y, \sum_S\left(w_k^2 - w_k\right)\sigma_{2k}^2$, that the last appearance of h_k on the right-hand side of equation (4.2) is not squared, as it would be if $h_k e_{2k}$ substituted for y_k everywhere. From a quasi-design viewpoint, h_k is asymptotically identical to unity, so whether or not it is squared makes no asymptotic difference.

Observe that the h_j' have been inserted in equation (4.3) for the same reason as α' was inserted into $\mathbf{b}$ in equation (3.1). Since the h_j' are asymptotically unity, however, they are not really needed (and serve no function whatever from a prediction-model viewpoint). A similar argument applies to the h_f in equation (4.4): they are asymptotically unity from the quasi-sampling-design viewpoint (and part of an estimate of 0 from a prediction-model viewpoint).

5 Some simulations

Paralleling Kott and Liao (2012), we generated a synthetic population, U, of hospitals from the 2008 DAWN public-use file. After creating U, we independently drew 3,600 stratified simple random samples of size 400 from U using the strata definitions on the public-use file. These definitions incorporate information on location and hospital ownership (public or private) not directly provided on the file.

We set the stratum sample sizes roughly proportional to a size measure q_k, but never less than four. For q_k we used annual drug-related emergency-room visits, which was always positive. The DAWN actually has a size variable attached to every hospital in the frame: total emergency-room visits in a previous year according to the American Hospital Association. Unfortunately, it was not included on the public-use file. Design weights in our simulations varied between 4.375 and 48, which allowed us to treat the finite population correction factors as ignorable in variance estimation.

As in our original paper, we generated a respondent sample R for each simulated sample based on Bernoulli draw from the logistic function:

$$p_k = (1 + \exp(3.735 - 0.4\log(q_k)))^{-1}, \tag{5.1}$$

We also created alternative respondent samples using

$$p_k = \left(1 + \exp\left(0.597 - 0.005q_k^{1/2}\right)\right)^{-1}. \tag{5.2}$$

Both response models produce unweighted overall response rates of around 54%, which is similar to actual DAWN experience, where response is also a mildly increasing function of the size variable. Notice that $\alpha_k = 1/p_k$ is bounded even if neither probability can be expressed by equation (2.4) with a finite u.

As in the previous study, we focused on estimating population totals for three survey variables. Annual drug-related emergency-room visits with adverse pharmaceutical reaction and those resulting in deaths came from the public-use file. Since both these variables were roughly linear in our size measure, the third "survey" variable was artificially constructed. It was the size measure (annual drug-related emergency-room visits) raised to the 1.3 power.

We investigated eight estimators and estimates of their variance. These are summarized in Table 5.1. The first two featured calibration to the original sample only (equation (2.5) with $\theta = 1$), with response assumed to be logistic in the log of the size measure. That is to say, equation (2.3) was employed with $\mathbf{x}_k = (1 \ \log(q_k))^T$. The first estimator used $\mathbf{z}_k = (1 \ \log(q_k))^T$ as the calibration vector while the second used $\mathbf{z}_k = (1 \ q_k)^T$, which was more consistent with a reasonable prediction model, at least for adverse reactions and deaths.

Our third and fourth estimator featured calibration to the sample and population in a single step (equation (2.5) with $\theta = 1$ and then $\theta = 0$) using $\mathbf{x}_k = \mathbf{z}_k = (1 \ \log(q_k) \ q_k)^T$. They were designed to be nearly unbiased if either the logistic response model in $(1 \ \log(q_k))^T$ or the linear prediction model in $(1 \ q_k)^T$ held.

Table 5.1
Summary of simulation exercise (all results in percentages %)

Estimator	t_{y1}	t_{y2}	t_{y3}	t_{y4}	t_{y5}	t_{y6}	t_{y7}	t_{y8}
Calibration to Sample								
response-model variables: $\mathbf{x}_{1k}$	$(1\ \log(q_k))^T$	$(1\ \log(q_k))^T$	$(1\ \log(q_k)\ q_k)^T$	-	$(1\ \log(q_k))^T$	$(1\ \log(q_k))^T$	$(1\ \log(q_k))^T$	$(1\ \log(q_k))^T$
calibration variables: $\mathbf{z}_{1k}$	$(1\ \log(q_k))^T$	$(1\ q_k)^T$	$(1\ \log(q_k)\ q_k)^T$	-	$(1\ \log(q_k))$	$(1\ q_k)^T$	$(1\ \log(q_k))^T$	$(1\ q_k)^T$
Calibration to Population								
response-model variables: $\mathbf{x}_{2k}$	-	-	-	$(1\ \log(q_k)\ q_k)^T$	$(1\ \log(q_k)\ q_k)^T$	$(1\ \log(q_k)\ q_k)^T$	$f_k\ (1\ \log(q_k)\ q_k)^T$	$f_k\ (1\ \log(q_k)\ q_k)^T$
calibration variables: $\mathbf{z}_{2k}$	-	-	-	$(1\ \log(q_k)\ q_k)^T$	$(1\ \log(q_k)\ q_k)^T$	$(1\ \log(q_k)\ q_k)^T$	$(1\ \log(q_k)\ q_k)^T$	$(1\ \log(q_k)\ q_k)^T$
True Response: $p_k = 1/\{1 + \exp[3.735 + 0.4 \log(q_k)]\}$								
Adverse Reactions								
Relative Bias of t_y	-0.07	0.06	-0.11	-0.13	-0.02	-0.07	0.10	0.09
Relative RMSE of t_y	4.97	3.98	4.01	2.45	2.51	2.57	2.40	2.39
Relative Bias of $v(t_y)$	8.60	12.59	12.52	6.24	6.76	6.16	6.76	6.48
Deaths								
Relative Bias of t_y	-0.17	0.06	-0.20	-0.26	-0.20	-0.30	0.04	-0.07
Relative RMSE of t_y	11.75	11.39	11.56	11.07	11.28	11.36	10.91	10.91
Relative Bias of $v(t_y)$	-1.34	-0.48	-0.90	-0.76	-1.00	-0.60	-0.12	-0.28
$(\text{Size})^{1.3}$								
Relative Bias of t_y	-0.16	-0.05	0.08	0.09	0.04	0.06	-0.02	0.01
Relative RMSE of t_y	6.92	5.07	5.06	0.95	1.05	1.12	0.89	0.89
Relative Bias of $v(t_y)$	10.01	18.49	17.47	-2.26	-3.41	-3.32	0.51	-2.12
True Response: $p_k = 1/\{1 + \exp[0.597 + 0.005\ q_k^{1/2}]\}$								
Adverse Reactions								
Relative Bias of t_y	2.87	-0.26	0.08	0.04	0.48	0.53	0.15	0.07
Relative RMSE of t_y	5.90	3.97	4.00	2.35	2.43	2.45	2.33	2.35
Relative Bias of $v(t_y)$	-18.22	11.63	11.95	9.90	8.82	7.35	7.19	6.67
Deaths								
Relative Bias of t_y	1.24	-1.88	0.47	0.36	1.03	1.20	-0.58	-0.67
Relative RMSE of t_y	11.42	11.01	11.41	10.95	11.18	11.26	10.69	10.72
Relative Bias of $v(t_y)$	5.30	3.00	6.27	6.24	5.65	5.06	6.21	5.90
$(\text{Size})^{1.3}$								
Relative Bias of t_y	5.17	1.05	-0.07	-0.05	-0.31	-0.36	0.01	0.08
Relative RMSE of t_y	9.11	5.31	5.05	0.85	0.97	1.01	0.80	0.82
Relative Bias of $v(t_y)$	-26.83	11.70	17.09	8.23	0.29	-3.98	5.17	2.90

$f_k = d_k \alpha_k - 1 = (d_k / \hat{p}_k) - 1$

Not surprisingly, the (empirical) relative mean squared error of the fourth estimator is always lower than the third. The reason is fairly obvious looking at equation (3.1) and considering the consequence of θ being 0 (calibration to the population) rather than 1 (calibration to the sample).

The fifth through eighth estimators were calibrated in two steps. The fifth and seventh estimators employed the calibration weighting from the first estimator in its first step, while the sixth and eighth employed the calibration weighting from the second estimator. The fifth and sixth used $\mathbf{z}_{2k} = \mathbf{x}_{2k} = (1 \ \log(q_k)\, q_k)^T$ in their second step, while the seventh and eighth were nearly pseudo-optimal (Kott 2011) using $\mathbf{z}_{2k} = (1 \ \log(q_k)\, q_k)^T$ and $\mathbf{x}_{2k} = (d_k\alpha_k - 1)\,\mathbf{z}_{2k}$ in their second step. All four employed the individual weight-adjustment functions:

$$h_k\left(\mathbf{g}_2^T\mathbf{x}_{2k}\right) = \frac{1}{d_k\alpha_k} + \left(1 - \frac{1}{d_k\alpha_k}\right)\exp\left[\frac{\mathbf{g}_2^T\mathbf{x}_{2k}}{1 - \dfrac{1}{d_k\alpha_k}}\right].$$

As Kott (2011) showed these $h_k\left(\mathbf{g}_2^T\mathbf{x}_{2k}\right)$ are asymptotically identical to the weight-adjustment function, $1 + \mathbf{g}_2^T\mathbf{x}_{2k}$, when $\mathbf{g}_2^T\mathbf{x}_{2k} = \mathrm{O}_P\left(1/\sqrt{n}\right)$ but prevent any w_k from falling below unity. Each is a version of equation (4.1) with $\ell_k = 1/(d_k\alpha_k)$, $c = 1$, and $u = \infty$.

Because the nonresponse rate was so large, we did not encounter a problem computing the third and fourth estimator using any of the simulated respondent samples. The relative mean squared error of the fourth estimator was always slightly higher than that of the seventh and eighth estimators, which incorporated nearly pseudo-optimal calibration in their second step. Interestingly, this was not the case when comparing the fourth estimator to the fifth and sixth estimators which, although employing two steps, did not incorporate nearly pseudo-optimal calibration.

Observe that although the second estimator always had a smaller relative mean squared error than the first, being more consistent with a reasonable prediction model (even for $q_k^{1.3}$, the survey variable appeared closer to being linear in q_k than in $\log(q_k)$), the other analogous pairs (fifth vs sixth and seventh vs eighth) exhibited no clear pattern of superiority. This is because it is the second-step residuals that are effectively modeled in equation (4.4) not the y-values.

Generating the nonresponse with equation (5.2) than (5.1) did not seem to have much of an impact on the results except for the relative biases of the first estimator. For both adverse reactions and $(\text{size})^{1.3}$, the relative bias of this estimator is over 40% of the relative mean squared error. That is likely because both models that could be used to justify this estimator (response is logistic in the log of the size measure and the survey variable is linear in the log of the size measure) fail. Not surprisingly, since the relative bias is such a large part of the relative mean squared error in these two situations, $v(t_k)$ underestimates mean squared error badly. Nowhere else is the relative bias of $v(t_k)$ greater than 15%.

It seems that even our artificial variable, $(\text{size})^{1.3}$, was close enough to being linear in the size measure that bias was never an issue for any estimator other than the first. The first estimator itself had a negligible relative bias when response was a logistic model of the log of the size measure, as assumed.

6 Concluding remarks

In Section 4, we noted two reasons to prefer calibration weighting in two steps: to make implicitly fitting a logistic response model easier and to incorporate nearly quasi-optimal calibration. A side benefit

of two-step calibration is more efficient estimation of the response model in step one since there is no sampling error to confound the estimation. This is useful when one wants to analyze the causes of unit nonresponse for its own sake.

We must concede, however, that the reduction in mean squared error using two steps was modest in our simulation experiments in Section 5. Moreover, the practical appeal of the simplicity of calibrating in a single step cannot be denied.

When calibration-weighting is used to adjust for nonresponse that is not missing at random as described in Chang and Kott (2008) and Kott and Chang (2010), the efficiency gains from a second step involving only calibration variables and functions of calibration variables model variables is likely to be sizeable.

When the finite population correction factors can be ignored, replication offers a much simpler approach to variance estimation than equation (3.7) even though the second summation on the right-hand side can be dropped in this situation. A different attractive alternative is the "collapsed" version of equation (4.2) that ignores the impact of the first calibration step:

$$\tilde{v}\left(t_y\right) = \sum_{k,j\in S}\left(1-\frac{\pi_k\pi_j}{\pi_{kj}}\right)\left[w_k e_{2k}\right]\left[w_j e_{2j}\right] + \sum_{k\in R} d_k\left(h_k^2\alpha_k^2 - h_k\alpha_k\right)e_{2k}^2.$$

This estimator clearly estimates the prediction-model variance if that model holds. A version of it – with the second summation removed – fared well in our simulation experiments (not shown). Some caution is needed before one draws too strong a conclusion from that result since the linear model was never too far from holding in our investigations.

Finally, a number of assumptions were made to simplify the exposition. The interested reader can extend the results to unbounded d_k, more general and not-necessarily-bounded weight-adjustment functions, or to allow the prediction-model errors to be correlated within primary sampling units. When N grows faster than n, the assumption that $\sigma_k^2 = \mathbf{z}_k^T\boldsymbol{\eta}$ can sometimes be dropped. See, for example, Kott (2009, page 69).

Acknowledgements

This paper was prepared for the Symposium on the Analysis of Survey Data and Small Area Estimation, in honour of the 75th Birthday of Professor J.N.K. Rao sponsored by the Fields Institute for Research in Mathematical Sciences. The authors would like to thank the organizers of the conference for the invitation to present this paper and the Institute for its generous funding of the conference without which this paper would never have been written. They would also like to thank a number of editors and referees for their helpful comments.

References

Bang, H., and Robins, J.M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61, 962-972.

Bankier, M. (2002). Regression estimators for the 2001 Canadian Census. Presented at the International Conference in Recent Advances in Survey Sampling.

Chang, T., and Kott, P.S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika,* 95, 557-571.

Deming, W.E., and Stephan, F.F. (1940). On a least squares adjustment of a sample frequency table when the expected marginal total are known. *Annals of Mathematical Statistics,* 11, 427-444.

Deville, J.-C. (2000). Generalized calibration and application to weighting for non-response. In *COMPSTAT: Proceedings in Computational Statistics, 14th Symposium, Utrecht, The Netherlands,* (Eds., J.G. Bethlehem and P.G.M. Van der Heidjen), Heidelberg: Physica Verlag, 65-76.

Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association,* 87, 418, 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.

Estevao, V.M., and Särndal, C.-E. (2002). The ten cases of auxiliary information for calibration in two-phase sampling. *Journal of Official Statistics,* 18, 233-255.

Estevao, V.M., and Särndal, C.-E. (2000). A functional form approach to calibration. *Journal of Official Statistics,* 16, 379-399.

Folsom, R.E. (1991). Exponential and logistic weight adjustments for sampling and nonresponse error reduction. *Proceedings of the American Statistical Association, Social Statistics Section,* 197-202.

Folsom, R.E., and Singh, A.C. (2000). The generalized exponential model for sampling weight calibration for extreme values, nonresponse, and poststratification. *Proceedings of the American Statistical Association, Survey Research Methods Section,* available online at http://www.amstat.org/sections/srms/Proceedings/, 598-603.

Fuller, W.A., Loughin, M.M. and Baker, H.D. (1994). Regression weighting in the presence of nonresponse with application to the 1987-88 Nationwide Food Consumption Survey. *Survey Methodology,* 20, 1, 75-85.

Kim, J.K., and Haziza, D. (2014). Doubly robust inference with missing survey data. *Statistica Sinica,* 24, 375-394.

Kim, J.K., and Park, H. (2006). Imputation using response probability. *Canadian Journal of Statistics,* 34, 1-12.

Kim, J.K., and Shao, J. (2013). *Statistical Methods for Handling Incomplete Data,* London: Chapman and Hall/CRC.

Kott, P.S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology,* 32, 2, 133-142.

Kott, P.S. (2009). Calibration weighting: Combining probability samples and linear prediction models. In *Handbook of Statistics 29B: Sample Surveys: Inference and Analysis,* (Eds., D. Pfeffermann and C.R. Rao), New York: Elsevier.

Kott, P.S. (2011). A nearly pseudo-optimal method for keeping calibration weights from falling below unity in the absence of nonresponse or frame errors. *Pakistan Journal of Statistics,* 27, 391-396.

Kott, P.S., and Chang, T.C. (2010). Using calibration weighting to adjust for nonignorable unit nonresponse. *Journal of the American Statistical Association*, 105, 1265-1275.

Kott, P.S., and Liao, D. (2012). Comparing weighting methods when adjusting for logistic unit Nonresponse. Presented at Federal Committee on Survey Methodology Research Conference, available online at http://www.fcsm.sites.usa.gov/files/2014/05/Kott_2012FCSM_III-B.pdf..

Little, R.J., and Rubin, D.B. (2002). *Statistical Analysis with Missing Data* (2nd Ed.), New York: John Wiley & Sons, Inc.

Lundström, S., and Särndal, C.-E. (1999). Calibration as a standard method for the treatment of nonresponse. *Journal of Official Statistics*, 15, 305-327.

Oh, H.L., and Scheuren, F.J. (1983). Weighting adjustment for unit nonresponse. In *Incomplete Data in Sample Surveys*, (Eds., W.G. Madow, I. Olkin and D.B. Rubin), New York: Academic Press, 2.

Rao, J.N.K. (1994). Estimation of totals and distributing functions using auxiliary information at the estimation stage. *Journal of Official Statistics*, 10, 153-165.

Robins J.M., Rotnitzky A. and Zhao L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89, p. 846-866.

Survey Methodology, June 2015
Vol. 41, No. 1, pp. 183-195
Statistics Canada, Catalogue No. 12-001-X

The relevance of follow ups in data collection for the Quality Assurance system of the Portuguese Population and Housing Census

Paula Vicente, Elizabeth Reis and Álvaro Rosa[1]

Abstract

The operationalization of the Population and Housing Census in Portugal is managed by a hierarchical structure in which Statistics Portugal is at the top and local government institutions at the bottom. When the Census takes place every ten years, local governments are asked to collaborate with Statistics Portugal in the execution and monitoring of the fieldwork operations at the local level. During the Pilot Test stage of the 2011 Census, local governments were asked for additional collaboration: to answer the Perception of Risk survey, whose aim was to gather information to design a quality assurance instrument that could be used to monitor the Census operations. The response rate of the survey was desired to be 100%, however, by the deadline of data collection nearly a quarter of local governments had not responded to the survey and thus a decision was made to make a follow up mailing. In this paper, we examine whether the same conclusions could have been reached from survey without follow ups as with them and evaluate the influence of follow ups on the conception of the quality assurance instrument. Comparison of responses on a set of perception variables revealed that local governments answering previous or after the follow up did not differ. However the configuration of the quality assurance instrument changed when including follow up responses.

Key Words: Quality assurance; Local government surveys; Follow ups; Map of Alert.

1 Introduction

The latest Portuguese Population and Housing Census took place in March 2011. It was a large and expensive statistical operation involving in-person, door-to-door contacts for the distribution and collection of paper questionnaires across the entire country. The foremost task of any census operation is to do a headcount of every person and identify where they live, without omitting anyone (Waite 2007). However, the successful accomplishment of such task can be compromised by various factors, notably the performance of the human resources involved, the level of citizens' cooperation and the specific characteristics of the regions and populations that are to be enumerated. Reliable data can only be obtained with sound and accurate processes, which is why the Census is assisted by a comprehensive Quality Assurance (QA) system that is designed and implemented throughout with the census operation itself (Wroth-Smith, Abbott, Compton and Benton 2011).

Prior to 2011, the QA system of Census operations was designed with standardized nationwide procedures i.e., standards, indicators, processes, and sub-processes were defined at national level and this also meant that all regions used the same QA activities for monitoring purposes. Although Portugal is a small country, it is geographically and demographically very diverse with heavily urbanized as well as rural areas; very densely populated areas and also villages that are almost abandoned and deserted; regions with predominately old people and other much younger regions. This diversity is likely to affect the implementation of a census operation as the problems, difficulties and risk of failure are not uniform, but

1. Paula Vicente, Elizabeth Reis and Álvaro Rosa, Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit (BRU-IUL), Lisbon, Portugal, ISCTE-IUL, Av. Forças Armadas, 1649-026 Lisbon, Portugal. E-mail: paula.vicente@iscte.pt.

vary according to the specificities of the population and the areas where the Census is being implemented. In light of this, a new perspective was introduced in the 2011 Census - the QA system was redesigned to tailor it to the local specificities of geographical areas and populations (Statistics Portugal 2007). This change required the Portuguese territory to be mapped in terms of risk of failure and the Map of Alert (Statistics Portugal 2010) was developed for this purpose.

Portugal is organized administratively into 303 municipalities encompassing 4,260 *freguesias* (local government units) (*freguesia* is the smallest administrative/governmental area in Portugal. Each Municipality comprises a set of *freguesias*. *Freguesia* is the equivalent to civil parish). This organization serves as the base for implementing Census operations: the *freguesia* is the lowest level of the operation's coordinating hierarchy; above it comes the municipal coordination, then the regional coordination and, finally, the national coordination is at the top. The Census Office of Statistics Portugal is in charge of the strategic and national coordination of the entire operation. Statistics Portugal appoints regional delegates for regional coordination; the presidents of municipalities are responsible for the municipal coordination, and finally the Presidents of *Junta de freguesia* (PJF) are in charge of the *freguesia* coordination. (*Junta de freguesia* is the governing body of each *freguesia*. The *Junta de freguesia* is administered by the President of *Junta de freguesia*).

The Map of Alert is a detailed map of the Portuguese territory at the *freguesia* level, in which each *freguesia* is attributed a colour code to indicate the potential risk of failure in the Census operation: red (high risk), orange (medium risk) and green (low risk). By risk of failure we mean possible coverage problems, i.e., failing to enumerate some population units or duplicating the enumeration of others. Mapping all the 4,260 *freguesias* according to their risk of failure would enable municipal coordinators to know in advance which *freguesias* would require specific QA activities in order to effectively assist the fieldwork operations. This would allow resources to be targeted to *freguesias* with a known high risk of non-accomplishment. Green or orange *freguesias* might therefore be treated with the standard QA procedures but specific procedures would be designed and implemented in line with local specificities of red *freguesias*. These might include assigning more experienced enumerators to the most difficult areas, controlling enumerators' work more regularly or checking more than the usual 5% of enumerators' work.

Information about characteristics of the populations, housing and areas that might cause coverage difficulties for the census (e.g., the existence of homeless people, of people belonging to minority groups or of areas with many vacant dwellings (Groves 1989, page 137, Groves and Couper 1998, page 176)) was necessary to build the Map of Alert. This kind of information could have been obtained from the 2001 Census data but, as this was potentially outdated, it was decided to collect the necessary information by means of a mail survey targeting all PJFs. It was crucial to get the cooperation of all 4,260 PJFs to ensure that each *freguesia* was classified with a risk level in the Map of Alert.

The questionnaire of the Perception of Risk survey was mailed at the beginning of October 2010. The deadline for data collection set internally by the research team was mid-December 2010 but, as respondents tend to postpone answering mail surveys, they were asked to send the completed questionnaires within one month. More than half of the *freguesias* (58%) returned the completed questionnaires within that time lag; after that period, responses continued to arrive but at a slower pace. By the deadline, 77% of the *freguesias* had returned the questionnaire but there was already a sharp decline in the number of questionnaires coming in at the end. Despite the good response rate (Dillman,

Smyth and Christian 2009), the goal of obtaining data from all the *freguesias* was far from reached. Ending the data collection in mid-December would have meant leaving nearly one fourth of the *freguesias* with no assigned risk level which would have drastically reduced the efficacy of the Map of Alert as an instrument of quality assurance. A follow up mailing to the non responding *freguesias* was therefore sent out on 16th December. Besides increasing sample size, it was expected that a qualitative gain would be obtained for the conception of the Map of Alert. In fact, there was some concern that non responders might be *freguesias* with problematic characteristics for the census, thus causing the true size of the red code to be underrepresented in the Map. The request for personal or sensitive information in questionnaires is known to increase the danger of nonresponse (e.g., Groves, Fowler Jr, Couper, Lepkowski, Singer and Tourangeau 2004, page 224) and although the information requested in the Perception of Risk survey was not personal (i.e., related to the PJF himself), it conveyed matters that the PJFs might be reluctant to share. Questions on the existence of homeless people, areas without public illumination, or roads without tarmac in the areas they govern might be considered overly sensitive thus leading to non participation in the survey. The follow up mailing also aimed to minimize the effect of non response on risk level classification of *freguesias*.

The Map of Alert was used in the Portuguese Population and Housing Census for the first time in its 2011 edition but Statistics Portugal intends to adopt it as a permanent QA instrument in future census operations. The study reported in this paper examines the impact of follow ups on the response rate and results of the Perception of Risk survey and evaluates to what extent the responses from follow ups changed the configuration of the Map of Alert, namely regarding the risk level classification.

The method used is presented in Section 2. Results are given in Section 3. Finally, a discussion is offered in Section 4.

2 Method

The Perception of Risk survey took place during the Pilot Test stage of the 2011 Portuguese Population and Housing Census (the Pilot Test was the last preparatory stage of the Census 2011 and took almost all of 2010). The aim was to collect information about any specific characteristics of *freguesias* that might hinder the exhaustive and accurate count of individuals and dwellings. The target population was defined as the *freguesias* of Portugal (N = 4,260). The Presidents of *Juntas de freguesia* were chosen to be the respondents because they have close contact with the populations and a deep knowledge of the problems of the areas they govern.

The questionnaire consisted of two blocks of questions (the questionnaire is presented in Figure A.1 of the Appendix). The first block included questions on the respondent's age, education, time as President of *Junta de freguesia*, frequency of computer and internet use, and the identification of the *freguesia* and municipality. The second block included questions on *freguesias* features potentially affecting the implementation of the census. This block had four sections. The first section contained a set of six items asking about characteristics of the *freguesia's* population. Respondents rated their answers on each of the items using a five-point scale ranging from "few" to "many". The second section contained a set of six items asking respondents about characteristics of the buildings and areas of the *freguesia*. Again each of the items was to be answered using a five-point scale ranging from "few" to "many". The next section

contained two items about enumerators' recruitment that were to be answered using a five-point scale ranging from "hard" to "easy". The questionnaire ends with one item on the overall perception about the implementation of the Census 2011 in the *freguesia*.

Statistics Portugal has an updated list of postal addresses of all *Juntas de freguesia* which was used as the sampling frame. The initial mailing was sent to all 4,260 PJFs, therefore making the Perception of Risk survey more of a census than a survey. The mailing included a questionnaire, a postage-paid return envelope and a cover letter. The letter and questionnaire were printed on paper with the logos of Census 2011 and Statistics Portugal responsible for implementing and coordinating the survey. Since responding to the survey was not compulsory, survey salience was emphasized in the invitation letter with the aim of improving the cooperation rate (e.g., Porter 2004, Dillman, et al. 2009): the letter explained that the survey concerned the Census 2011 operation and the PJFs' answers would be indispensable to the quality of the operation at both the local and national levels. Moreover, the importance of the response was underlined by the fact that the request came from Statistics Portugal.

All *freguesias* that had not returned the questionnaire by 15th December 2010 were sent the follow up mailing containing a second copy of the questionnaire, a cover letter insisting on response and a postage-paid return envelope. Data collection came to an end in mid-February 2011.

3 Results

For the purpose of the analysis, we shall consider two "groups" of responses: the initial group and the final group. The initial group includes the *freguesias* that returned the questionnaires before the follow up date; the final group includes all the *freguesias* responding to the survey, i.e., the initial group plus the *freguesias* that returned the questionnaires after the follow up. The two groups are not mutually exclusive.

The analysis starts with a description of the mailing outcomes. We examine response rates (overall and by region) and geographical distribution of the *freguesias* that could be assigned a risk level (both in the initial and the final versions of the Map of Alert). When making analyses by region, we use the NUTS II classification of the Portuguese territory; this entails six regions - North, Center, Lisbon, Alentejo, Algarve and Archipelagos of Madeira and Azores. In the second stage of the analysis, the responses of the PJFs are analysed by means of Principal Component Analysis with the purpose of reducing the dimensionality of the data and identify latent dimensions of risk. This analysis is performed in both groups of response. Finally an evaluation of the *freguesias'* risk level classification is made in both the initial and final Map of Alert. *Freguesias* that did not respond at all to the Perception of Risk survey (referred as non responders) are described according to their geographical distribution.

3.1 Analysis of response rates

Figure 3.1 presents the distribution of the number of questionnaires received per day during the overall collection period (from 10th October 2010 when the first questionnaires were received until the final deadline on 16th February 2011). There are two peaks of response, the first approximately one month after the first mailing went out and the second some days after the follow up mailing. Almost no questionnaires were being received by the time the follow up mailing was sent out, which leads us to believe that no more would have been received without the second mailing.

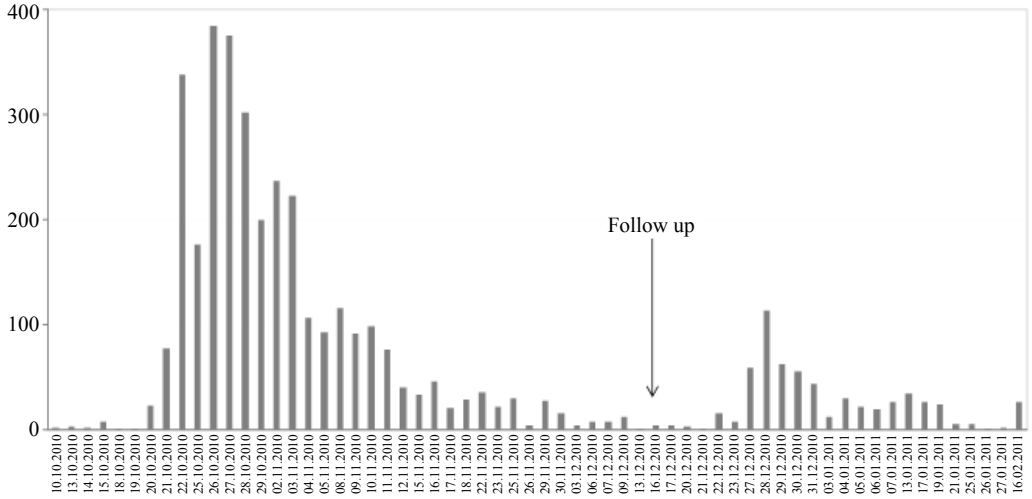


Figure 3.1 Number of questionnaires received per day

From a total of 4,260 questionnaires sent in the first mailing, 2,457 were answered within the suggested time of response (one month), 816 were answered after that period but before the follow up mailing and 609 were answered after the follow up date. Of the 4,260 *freguesias*, 378 did not respond. This absence of response was considered to be a refusal since it is unlikely these questionnaires were not delivered as an updated address list was used for mailing. The overall response rate of the survey, computed as the percentage of *freguesias* that answered the questionnaire out of the total number of *freguesias* in the population, was 91.1% (Table 3.1).

Table 3.1
Outcomes of the mailing of the questionnaires

	N	%
Freguesias returning the questionnaire within one month	2,457	57.7
Freguesias returning the questionnaire after one month and before the follow up mailing	816	19.2
Freguesias returning the questionnaire after the follow up mailing	609	14.3
Freguesias not returning the questionnaire	378	8.9
Questionnaires sent	4,260	100.0
Overall *freguesias* returning the questionnaire	3,882	91.1

Table 3.2 presents the response rate per region in the initial and final group. The response rate of the initial mailing ranged from 71% in the North to 88.1% in the Algarve; the final response rate ranged from 87.3% in the North to 96.4% in the Algarve. The follow up mailing allowed an increase both in the overall response rate and in the response rate of each region, but it was more efficient in the North than in other regions. The North had a 16.3% increase in survey participation, in contrast to an increase of approximately 6% in the region of the Archipelagos of Madeira and Azores.

Table 3.2
Response rate per region by response group (%)

Region	Initial	Final
North	71.0	87.3
Center	79.3	91.4
Lisbon	84.3	95.3
Alentejo	83.1	96.0
Algarve	88.1	96.4
Archipelagos of Madeira and Azores	86.7	93.3
Overall	76.8	91.1

Table 3.3 presents the geographical distribution of the *freguesias* with an assigned risk level in the initial and in the final Map of Alert. More than 40% of the *freguesias* are located in the North and approximately 26% are located in the Center. When comparing the final distribution with that of all the *freguesias* in the population, the biggest differences are found in the regions of Lisbon (13.1% vs. 7.0%, meaning that the region of Lisbon is overrepresented in the Map of Alert) and Center (26.1% vs. 30.6%, meaning that the region of Center is underrepresented in the Map of Alert). The geographical distribution of the *freguesias* with an assigned risk level in the final Map is very similar to that of the initial Map.

As to the non responding *freguesias,* more than half are located in the North and approximately one fourth are located in the Centre. The other regions have less than 10% of the *freguesias* with no risk level assigned. This pattern is evident in both the initial and final group.

Table 3.3
Geographical distribution of *freguesias* with risk level assigned and non-responders in the Map of Alert by response group and *freguesias* in the population (%)

Region	*Freguesias* with risk level assigned		Non responders		Population
	Initial	Final	Initial	Final	
North	44.0	46.1	59.5	63.2	46.6
Center	26.7	26.1	23.1	23.8	30.6
Lisbon	13.7	13.1	8.4	6.1	7.0
Alentejo	7.7	7.5	5.2	2.6	8.9
Algarve	2.3	2.1	1.0	0.9	2.0
Archipelagos of Madeira and Azores	5.6	5.1	2.8	3.4	4.9
N =	3,264†	3,873†	987	378	4,260

† Nine *freguesias* of the initial group could not be assigned a risk level because the question on *freguesia* identification was not answered.

3.2 Analysis of the PJF responses

In order to simplify the structure of the survey data and identify the potential dimensions of risk affecting the Census operation, two Principal Component Analysis (PCA) were conducted. One of the PCA was made using the five questions about the characteristics of the PJF (age, educational level, time as president of *Junta de freguesia*, frequency of computer use and frequency of internet use); the other PCA was made using the Likert-type questions about *freguesias'* characteristics and enumerators recruitment (Sections 1 to 3 of the questionnaire). The eigenvalue over one criterion was adopted to extract the

components. Table 3.4 presents the number of principal components (PC) and the percentage of total variance they explain, based on varimax rotation. Both PCAs were performed in the initial and final group of *freguesias*.

The outcomes reveal that the responses obtained from initial *freguesias* have an identical structure in the latent dimensions of risk to the responses of the final group of *freguesias*. The sampling adequacy indicator for the PCA on PJF characteristics was reasonably good (KMO > 0.6) in both the initial and final *freguesias* data sets. In both data sets two principal components were extracted accounting for approximately 77% of the data variance. The PCs were named as: PC_A – PJF' skills and PC_B – PJF' experience.

Table 3.4
Characteristics of Principal Component Analyses by response group

Analysis characteristic	Initial	Final
PCA on PJF characteristics		
Kaiser-Meyer-Olkin measure of sample adequacy	0.687	0.685
PCs extracted	2	2
Variance explained	77.3%	77.2%
PCA on *freguesias'* characteristics		
Kaiser-Meyer-Olkin measure of sample adequacy	0.693	0.696
PCs extracted	5	5
Variance explained	61.4%	61.3%

The sampling adequacy indicator for the PCA on Likert-type questions was also reasonably good (KMO > 0.6) in both data sets. Five PCs were extracted, both in the initial and final data sets, accounting for nearly 61% of the data variance, namely: PC_1 – Hard to reach population, PC_2 – Enumerators with suitable skills and available to work in the census, PC_3 – Elderly population, PC_4 – Deserted areas and PC_5 – Areas with high vacancy rates for habitable housing.

Regarding the overall opinion about the degree of difficulty in implementing the Census 2011 operation (question on Section 4 of the questionnaire), the response of nearly 2/3 of the respondents was above the middle point of the scale in both response groups. In the initial group, 67.8% of the respondents rated their answers as level "4" or "5" on the response scale compared with 67.5% in the final group (Table 3.5).

Table 3.5
Overall opinion about the Census by response group (%)

	Initial	Final
1 – "hard"	1.7	1.7
2	3.8	3.8
3	26.7	27.0
4	38.4	37.9
5 – "easy"	29.4	29.6

3.3 *Freguesias'* risk level classification

The seven dimensions of risk found with both PCAs were then used as an input in Finite Mixture Modeling and Cluster Analysis to produce a segmentation of the *freguesias* (details and outputs of this

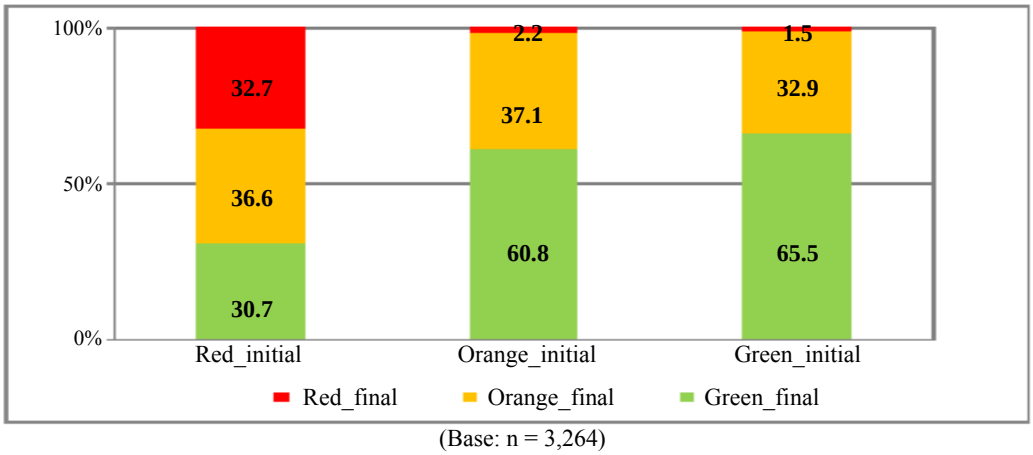
analysis are not presented but can be found on ISCTE-IUL (2011)). The segmentation is made for both the initial and final groups of *freguesias*. The outcome of the segmentation is presented in the Map of Alert in which the *freguesias* appear in red, orange or green (the final Map of Alert is presented in Figure A.2 of the Appendix. The dark spots represent the *freguesias* without an assigned risk level due to non response). Table 3.6 summarizes the *freguesias'* risk level classification in the initial and final versions of the Map.

Table 3.6
Risk level classification in the Map of Alert by response group (%)

Risk level	Initial (n = 3,264)	Final (n = 3,873)	Δ%
High risk (red)	6.4	3.7	– 42.2
Medium risk (orange)	53.3	33.9	– 36.4
Low risk (green)	40.3	62.4	+ 54.8

The dominant colour in the initial Map of Alert is orange (53.3% of the *freguesias* are rated as medium risk). The share of high risk *freguesias* is only 6.4%. Green predominates in the final Map (62.4% of the *freguesias* are classified as low risk) and less than 4% of the *freguesias* are red. Adding the follow up responses to the initial responses resulted in a change in the configuration of the Map of Alert, most notably the increase in the percentage of *freguesias* rated as low risk (+ 54.8%).

We then analysed how the follow up responses changed the risk level classification of the initial *freguesias*. The responses of the 3,264 initial *freguesias* allowed a colour code to be assigned to each *freguesia* and to draw the initial version of the Map of Alert. After incorporating the responses of the follow up *freguesias* the Map of Alert was redesigned – not only a higher number of *freguesias* could have a colour code assigned but also the colour initially attributed to the initial *freguesias* changed in some cases. Of the 3,264 initial *freguesias* approximately 50% got a different colour in the final Map of Alert. Figure 3.2 presents the overall changes in risk level classification of initial *freguesias* after integrating the responses of follow up *freguesias*.



(Base: n = 3,264)

Figure 3.2 Risk level classification in the final Map of Alert by risk level classification in the initial Map of Alert

The *freguesias* that were rated green at the outset (green_initial) tend to stay green (green_final) after considering the follow up responses (65.5%). Only 32.9% of the initially green *freguesias* changed to orange alert (orange_final) and 1.5% changed to red alert (red_final). As to the *freguesias* that started out as orange (orange_initial), the follow up responses caused 60.8% to change to green (green_final); only 37.1% remained orange (orange_final) and a minority of 2.2% changed to red (red_final). The biggest change caused by follow up responses is in the red group of *freguesias*: only 32.7% of the initially red *freguesias* (red_initial) remained high risk (red_final), and the majority changed either to orange (36.6%) or green (30.7%).

Finally, we analysed risk level classification per region, and compared the initial and final Map (Table 3.7).

Table 3.7
Risk level classification per region by response group (%)

Region	Risk level	Initial	Final	Δ%
North	High risk	4.2	0.8	–81.0
	Medium risk	52.7	46.1	–12.5
	Low risk	43.1	53.1	+23.2
Center	High risk	3.7	0.3	–91.9
	Medium risk	54.8	18.6	–66.1
	Low risk	41.5	81.2	+95.7
Lisbon	High risk	19.1	20.3	+6.3
	Medium risk	45.3	24.1	–46.8
	Low risk	35.6	55.6	+56.2
Alentejo	High risk	4.0	1.0	–75.0
	Medium risk	65.9	5.0	–92.4
	Low risk	30.1	94.0	+212.3
Algarve	High risk	17.0	29.8	+75.3
	Medium risk	43.4	38.1	–12.2
	Low risk	39.6	32.1	–18.9
A. Madeira and Azores	High risk	5.2	1.5	–71.2
	Medium risk	56.0	60.9	+8.8
	Low risk	38.8	37.6	–3.1

Lisbon and Algarve are the regions with higher percentage of *freguesias* coded as red (19.1% and 17.0%, respectively). This tendency holds both in the initial and final Map of Alert. The follow ups caused a reduction in the percentage of *freguesias* coded as red in all regions with the exception of Lisbon and Algarve for which the final Map of Alert presents higher percentages of red *freguesias* than the initial Map. Regarding the percentage of low risk *freguesias*, the follow ups caused and increased in all regions except Algarve and the Archipelagos of Madeira and Azores in which a decrease was noticed. Additionally, the percentage of orange *freguesias* decreased in all regions after adding the follow-up responses, except for the Archipelagos of Madeira and Azores.

4 Discussion

It is clear from the results shown above that the follow up mailing was valuable and had a positive impact on both the Perception of Risk survey response rate and the designing of the Map of Alert.

Although it was not possible to meet the 100% response rate target for the Perception of Risk survey, the high response rate – 91.1% - was only achieved thanks to the follow up mailing. The response rate differed across regions but the follow up allowed the response rate to go up in all the regions. North had the lowest response rate for both the initial group – 71% – and after the follow up – 87.3%. Several factors may have accounted for this result. Firstly, is the fact that the PJFs in the North remain in office longer than anywhere else in the country. The average time as PJF is 8.6 years in the North compared with a country average of 7.8 years. Moreover, whereas the 90th percentile of the "time as president" distribution is 20 years in the North, it does not exceed 17 years in the other regions. This means that the PJFs in the North have more governance experience and are probably able to make a better assessment of the impact of their *freguesias'* specificities on the census. Perhaps, these PJFs felt their *freguesias* would not present problems for the Census, so did not bother to answer the questionnaire. Another fact that might have accounted for the lower response rate in the North is that the main opposition party got the most votes in the North in the last parliament election so that the PJFs' lack of cooperation could have been a form of censure against the central government because they knew the survey had been requested by the country's official bureau of statistics. Finally, the North is the region with the most *freguesias* – nearly 2,000 – which makes a 100% response rate more difficult to achieve than in smaller regions like the Algarve, which has less than 90 *freguesias*.

Follow up responses led to changes in the risk level classification of the *freguesias*. Contrary to initial expectations, the scenario of color code in the final Map of Alert was not more problematic than the scenario in the initial Map. Not only was the percentage of red alert *freguesias* smaller in the final Map, but the percentage of green code *freguesias* also went up. Therefore, in addition to increasing the number of *freguesias* on the Map with an assigned risk level (from 3,264 *freguesias* to 3,873 *freguesias*) the follow up mailing also allowed the classification of some *freguesias'* to be "corrected", namely those initially classified as high risk, most of which were re-coded to orange or green after considering the data set from the follow ups.

These outcomes underline the importance of local governments being more involved and participating actively in future editions of the survey. The contact strategy adopted for the Perception of Risk survey was to send and receive the questionnaire by mail, but different approaches may be considered in the future, namely to include other modes such as the internet. Additionally, contact strategies could be customized to regions specificities. As the North had the lowest response rate, a strategy that included more follow up contacts (using the mail, the telephone or the e-mail) could be adopted there, and a less aggressive contact and re-contact strategy used in other regions. Finally, it must be noted that the administrative map of Portugal was changed in 2013 and the total number of *freguesias* has now been reduced to approximately 3,000. This new format of organization will surely favor the next Perception of Risk survey since a smaller number of PJFs will simplify the implementation of a contact strategy and the exhaustive inquiry of the *freguesias*.

Acknowledgements

This article is part of the project *Programa de Controlo e Avaliação da Qualidade dos Censos 2011*, a joint project of Statistics Portugal and *Instituto Universitário de Lisboa* (ISCTE-IUL).

Appendix

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

CENSOS 2011

Perception of risk survey

Questionnaire to Presidents of *Juntas de freguesia* as part of the Pilot Test of the 2011 Census

IDENTIFICATION

Freguesia: _____

Municipality:_____

Name: _____ Age: _____

Educational level:

Less than basic level ☐ Basic level (9 years compulsory) ☐ Secondary ☐ University ☐

For how long have you been president in this *Junta de freguesia*: _____ years

Frequency of computer use: Rarely ☐ Several times a day ☐ Several times a week ☐ Everyday ☐

Frequency of internet use: Rarely ☐ Several times a day ☐ Several times a week ☐ Everyday ☐

PERCEPTION ABOUT *FREGUESIAS'S* FEATURES

Rate your responses using a 1 to 5 scale for the following items regarding the *Freguesia*. Mark the number corresponding to your choice with X.

1 POPULATION							
1. Existence of elderly population (age ≥65 years)	Few	1	2	3	4	5	Many
2. Existence of illiterate population (cannot read or write)	Few	1	2	3	4	5	Many
3. Existence of population living in social housing neighbourhoods	Few	1	2	3	4	5	Many
4. Existence of emigrant population	Few	1	2	3	4	5	Many
5. Existence of immigrant population	Few	1	2	3	4	5	Many
6. Existence of homeless population	Few	1	2	3	4	5	Many
2 HOUSING AND AREAS							
1. Existence of areas with predominantly closed condominiums	Few	1	2	3	4	5	Many
2. Existence of areas with predominantly second or summer homes	Few	1	2	3	4	5	Many
3. Existence of areas with predominantly recently built residential housing	Few	1	2	3	4	5	Many
4. Existence of areas with difficult access (e.g., no tarmac roads, no lighting, …)	Few	1	2	3	4	5	Many
5. Existence of areas with dispersed housing	Few	1	2	3	4	5	Many
6. Existence of predominantly dormitory areas	Few	1	2	3	4	5	Many
3 HUMAN RESOURCES							
1. How difficult will it be to recruit suitably skilled enumerators	Hard	1	2	3	4	5	Easy
2. How difficult will it be to recruit enumerators with availability	Hard	1	2	3	4	5	Easy
4 OVERALL OPINION ABOUT THE CENSUS							
How difficult will it be to implement the Census 2011 operation in the *freguesia*	Hard	1	2	3	4	5	Easy

Figure A.1 Perception of risk questionnaire

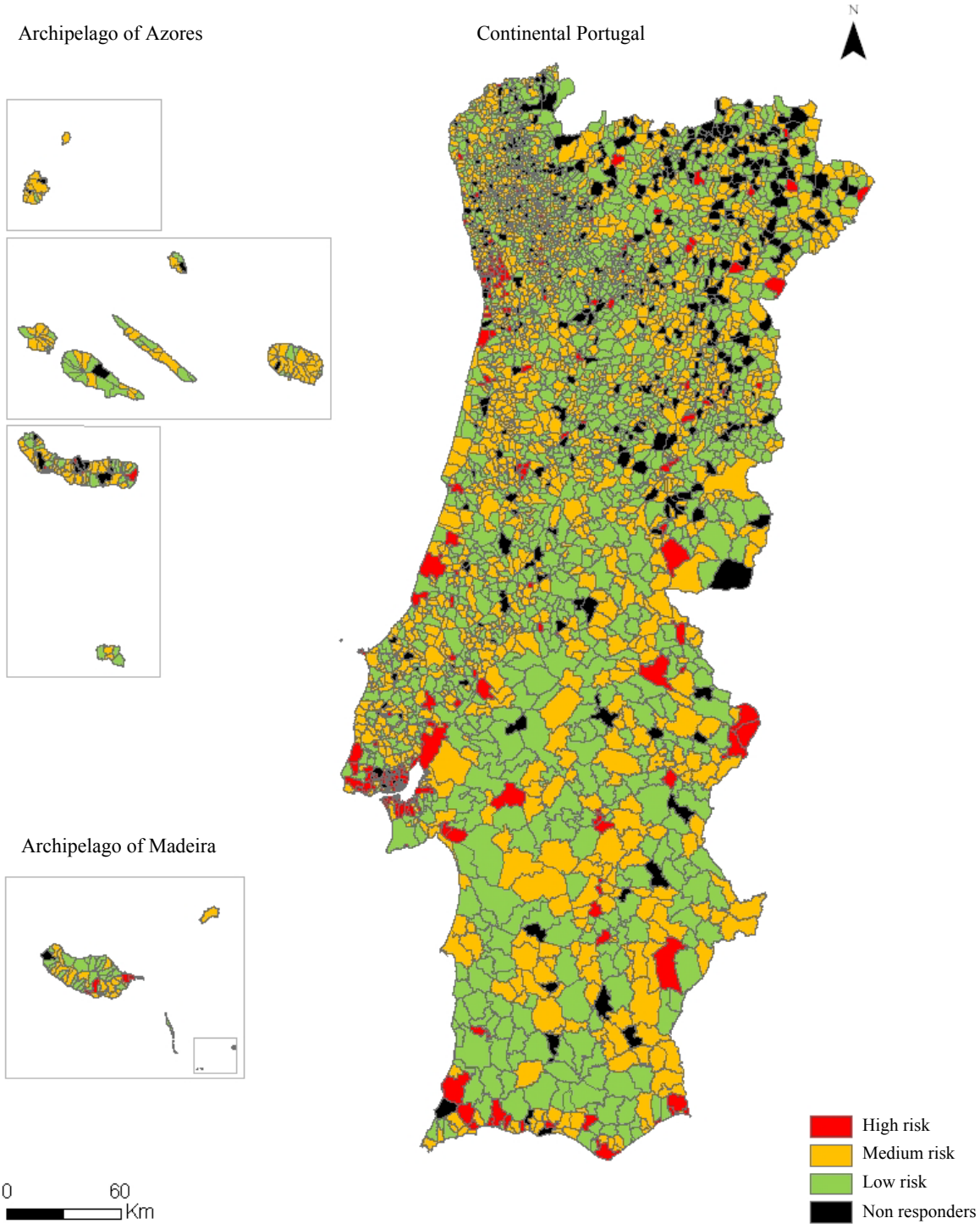


Figure A.2 Final Map of Alert

References

Dillman, D., Smyth, J. and Christian, L. (2009). *Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method*, 3rd Edition. New Jersey: Wiley.

Groves, R. (1989). *Survey Errors and Survey Costs*. New York: Wiley-Interscience.

Groves, R., and Couper, M. (1998). *Non-response in Household Interview Surveys*. New York: Wiley-Interscience.

Groves, R., Fowler Jr, F., Couper, M., Lepkowski, J., Singer, E. and Tourangeau, R. (2004). *Survey Methodology*. New York: Wiley-Interscience.

ISCTE-IUL (2011). Censos 2011-sistema de indicadores de alerta. (Unpublished document).

Porter, S. (2004). Raising response rates: What works? *New Directions for Institutional Research*, 121, 5-21.

Statistics Portugal (2007). *Programa de Acção para os Censos 2011*. Census Office, Statistics Portugal.

Statistics Portugal (2010). *Plano de Controlo e Avaliação da Qualidade Censos 2011 – Controlo do Processo Produtivo*. Census Office, Statistics Portugal.

Waite, P. (2007). *State, Local and Tribal Governments Benefit by Early Participation in the 2010 Census*. US Census Bureau Press Release.

Wroth-Smith, J., Abbott, O., Compton, G. and Benton, P. (2011). Quality assuring the 2011 Census population estimates. *Population Trends,* 143, 13-21.

Survey Methodology, June 2015
Vol. 41, No. 1, pp. 197-214
Statistics Canada, Catalogue No. 12-001-X

Measuring temporary employment. Do survey or register data tell the truth?

Dimitris Pavlopoulos and Jeroen K. Vermunt[1]

Abstract

One of the main variables in the Dutch Labour Force Survey is the variable measuring whether a respondent has a permanent or a temporary job. The aim of our study is to determine the measurement error in this variable by matching the information obtained by the longitudinal part of this survey with unique register data from the Dutch Institute for Employee Insurance. Contrary to previous approaches confronting such datasets, we take into account that also register data are not error-free and that measurement error in these data is likely to be correlated over time. More specifically, we propose the estimation of the measurement error in these two sources using an extended hidden Markov model with two observed indicators for the type of contract. Our results indicate that none of the two sources should be considered as error-free. For both indicators, we find that workers in temporary contracts are often misclassified as having a permanent contract. Particularly for the register data, we find that measurement errors are strongly autocorrelated, as, if made, they tend to repeat themselves. In contrast, when the register is correct, the probability of an error at the next time period is almost zero. Finally, we find that temporary contracts are more widespread than the Labour Force Survey suggests, while transition rates between temporary to permanent contracts are much less common than both datasets suggest.

Key Words: Temporary contracts; Measurement error; Hidden Markov model; Register data.

1 Introduction

The issue of temporary employment is receiving increased attention in the economic and political debate. Temporary contracts allow employers to circumvent strict hiring and firing regulations (Bentolila and Bertola 1990; Booth 1997; Cahuc and Postel-Vinay 2002) and some times even regulations concerning wage rigidity (OEDC 2002). Especially during economic recessions, temporary contracts are used by employers to adjust their labour force for product demand fluctuations.

The Netherlands has been a pioneer in flexible employment since the beginning of the 1990's. Contractual flexibility is an important feature of the Dutch labour market. Temporary employment rose sharply from 5.9% in 1991 to 17.1% in 2010 (OECD 2012), while growth in temporary employment contributed 9.9 percentage points to the total employment growth from 1990 to 2000 (OECD 2002). Employers have typically a "minimum capacity" personnel strategy (Sels and Van Hootegem 2001), meaning that companies employ their "core" workers with permanent contracts and offer temporary contracts to the rest to be able to adjust in times of an economic slump.

Whereas, in the Netherlands, statistics on temporary contracts were until recently based exclusively on data from household and labour force surveys, high-quality register data has become available that may be used in conjunction with – or even replace – the survey data. The first confrontation of the two data sources revealed some severely diverging figures in the size of temporary employment. In 2009, the share of all types of temporary contracts was 15.4% according to the Labour Force Survey (LFS), while 23.6% according to the "Polisadministratie" (PA) data, which are register data provided by the Institute for

1. Dimitris Pavlopoulos, VU University Amsterdam, Department of Sociology, De Boelelaan 1081, 1081 HV Amsterdam, the Netherlands. E-mail: d.pavlopoulos@vu.nl; Jeroen K. Vermunt, Tilburg University, Department of Methodology and Statistics, PO Box 90153, 5000 LE Tilburg, the Netherlands. E-mail: j.k.vermunt@tilburguniversity.edu.

Employee Insurance (UWV) (Hilbers, Houwing and Kösters 2011). As the size of temporary employment is very important for the design of labour market policies, Statistics Netherlands undertook the task of resolving the discrepancies between the two data sources. The results of the further investigation of the data were not very promising. Preliminary results indicate that 15.6% of those having a permanent contract according to the LFS appear to have a temporary contract according to the PA, while 18.3% of those having a temporary contract with duration shorter than one year according to the LFS appear to have a permanent contract according to the PA (Mars 2011). Although part of the inconsistencies can be explained by the somewhat different definitions of temporary employment in the two data sources, large discrepancies remain even when both using a matched sample and selecting the cases where no definitional differences exist.

As previous research suggests, measurement error can account for the encountered inconsistencies between the survey and register data. As far as survey data are concerned, measurement error has been recognized as an important source of bias (Rodgers, Brown and Duncan 1993; Pischke 1995; Bollinger 1996; Rendtel, Langeheine and Berntsen 1998; Bound, Brown and Mathiowetz 2001; Biemer 2011). Although no research exists on the error in the measurement of the contract type, research on other labour market characteristics, such as employment participation, wages, working hours, industry and occupation, indicates that survey data may contain large amounts of measurement error, which may severely bias the results of statistical analyses. For example, Biemer (2004) suggests that in the surveys of 1992-1994 of the Current Population Survey, 20.9% of the unemployed respondents were incorrectly classified to other states. Gottschalk (2005) indicates that two-thirds of the observed nominal-wage reductions without a job change were due to measurement error. Specifically, 17% of the workers report a nominal wage reduction from year to year while remaining with the same employer. However, when controlling for measurement error, yearly nominal wage reductions are faced by no more than 4-5% of the workers that remain with the same employer. Using the Panel Study of Income Dynamics (PSID) validation study, Mathiowetz (1992) suggests that company registers and survey responses in occupational classification agreed by 87.3%. Brown and Medoff (1996) find a 0.82 correlation of company registers and survey responses on the establishment size and a 0.86 on company size.

Research on measurement error in register data is clearly scarcer than on survey data. Register data are typically treated as error free and are used as a “golden standard” when confronted with survey data. For example, most research using the PSID validation study relies on this assumption (Duncan and Hill 1985; Rodgers et al. 1993; Bound, Brown, Duncan and Rodgers 1994; Pischke 1995). However, there is also research showing that the “golden standard” assumption may not be always plausible. Kapteyn and Ypma (2007) study measurement error in earnings and, although they retain the assumption that register data are error-free, they allow for errors in the matching of survey with register data. Specifically, they assume that a record in the register is identical to a record in the survey with a certain probability. They conclude that introducing this extra source of error changes the pattern of the measurement error in the survey. Abowd and Stinson (2005) compare earnings’ reports from the Survey of Income and Program Participation (SIPP) and the Detailed Earnings Records (DER). Measurement error is found to be larger in the administrative DER data (20%-27%) than in the SIPP data (13%-15%). Comparing the same data sets, Gottschalk and Huynh (2010) suggest that measurement error can severely bias measures of income inequality.

The aim of the current paper is to estimate the amount of error in the measurement of contract type in the Dutch LFS. For this purpose, the survey data are matched with register data from the PA. The register data are not treated as error-free, as we model simultaneously the measurement error in both sources. We

use an extended hidden Markov model with two indicators for the type of contract (temporary or permanent), each coming from one of our data sources.

The rest of the paper is organized as follows: in Section 2, we elaborate further on the problem of the measurement of temporary employment in the Netherlands by presenting the relevant details on the two data sources and showing some descriptive statistics. In Section 3, we present the hidden Markov model that was used in this study. Section 4 discusses the results of our analysis. The conclusions of our study are presented in Section 5.

2 Description of the two data sources

The two data sources providing information on temporary contracts are the Labour Force Survey (in Dutch: *Enquête Beroepsbevolking*) administered by Statistics Netherlands (in Dutch: *Centraal Bureau voor de Statistiek* – CBS) and the "Polisadministratie"-dataset of the Institute for Employee Insurance (UWV). The LFS is a rotating trimonthly survey on individual labour-market characteristics that is representative for the Dutch population older than 15 years of age. The survey was launched in 1987, while its longitudinal component was introduced in 1999. Since 1999, respondents are interviewed at five consecutive panel waves, which makes it possible to study short-term individual developments in the labour market. The information that is collected refers to the moment of the interview. The interviews are spread rather evenly within the trimester.

Errors in the measurement of the contract type in the LFS are, as is typical in surveys, the result of misreporting by respondents or mistakes in the recording of responses by interviewers. An additional error source is the use of proxy interviews. Typically, in the LFS, a single household member provides responses for all household members included in the sample, which increases the measurement error. In our LFS-sample, 40.1% of all observations refer to proxy interviews. A further possible cause of measurement error is that workers may confuse the legal employment contract with the implicit or psychological contract with their employer. Especially in younger cohorts where flexible contracts are widespread and in sectors with large job mobility and changing employment conditions, such as the health sector, workers may report that they have a permanent contract based on promises of the employer, while in reality they are employed on a temporary contract.

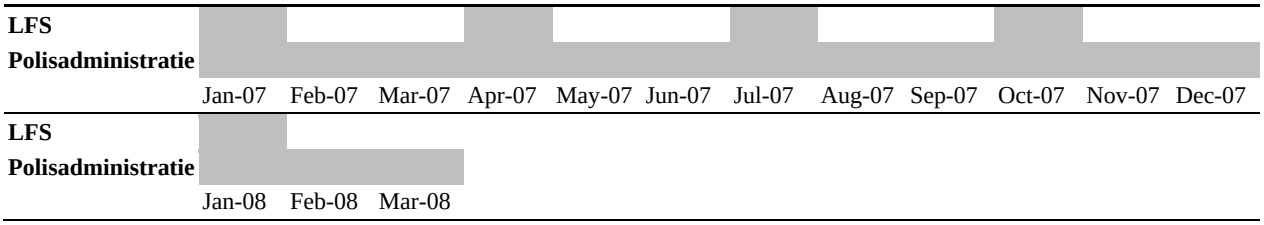
The PA is a unique register dataset containing labour market and income information for all insured workers in the Netherlands. This dataset is constructed by collecting and matching information from various sources, such as the Tax Office (in Dutch: *Belastingdienst*) – including data from individual tax-reporting statements (in Dutch: *jaaropgave*), declarations from temporary work agencies (in Dutch: *weekaanleveringen*) and the Population Register (in Dutch: *Gemeentelijke BasisAdministratie persoonsgegevens* – GBA). The PA is administered by the Dutch Institute for Employee Insurance (UWV).

The UWV has a strong interest in maintaining the high quality and accuracy of the PA as this data source is used by several governmental institutions. For example, the social security contributions, the housing allowance (in Dutch: *huurtoeslag*), and the health care allowance (in Dutch: *zorgtoeslag*) are determined using information from this dataset. To improve the data quality, the PA has undergone several revisions since 2006. There is no missing data as the submission of tax-reporting statements is compulsory for employers. However, whereas the dataset contains monthly information, employers

typically submit the relevant information only once per year (the moment of submission is not possible to be retrieved). This may create possible mistakes for the period between two consecutive submissions, especially in the measurement of the type of contract, which is clearly not the most important variable for the users of the PA. Therefore, we may expect that if a mistake is made in the contract type, it persists till the moment that the employer submits the following report to the UWV. This means that the measurement error in the PA can be expected to be serially correlated.

For our study, we select the LFS-respondents that were interviewed for the first time in the first trimester of 2007. Since we focus on employed individuals, we retained in the sample individuals aged from 25 to 55. After implementing the age restriction, we ended up with a sample size of 11,632 individuals. For all these individuals, the information from the LFS was matched with the monthly information from the PA by Statistics Netherlands using the social security number of individuals. The achieved matching level was 98% and all relevant inconsistencies were resolved (the matching and the quality control was done by Statistics Netherlands). Our final dataset has the form of a person-month file for 11,632 individuals with 15 observations corresponding to the period January 2007 – March 2008 and containing full information from the PA and partially observed information (5 observations – one response per 3 months) from the LFS. The matched dataset is illustrated in Table 2.1. This panel dataset is unbalanced for the LFS as our survey data suffer from some attrition. More specifically, from the 11,632 individuals that responded to the first interview, 9,970 were left in the LFS-sample in the second interview, 9,113 for the third, 8,953 for the fourth and 8,629 for the last interview. In the PA-data for this sample there is no attrition, so the sample is fully balanced.

Table 2.1
An illustration of our sample



Note: This illustrates how the rotation panel of the LFS corresponds to monthly observations from the Polisadministratie. This table refers to individuals that were interviewed every first month of the trimester. A cell that is shaded gray indicates a valid observation.

The variable of main interest for our study is the contract type, which takes on three possible values: permanent contract, temporary contract, and “other”.

The contract type is derived from the main job, which means that information on other jobs that individuals may hold is ignored. Individuals who are not in paid employment are classified as belonging to the “other” state. It should be noted that the latter state is rather heterogeneous as it includes among others the categories self-employed, unemployed, and in full-time education. However, the inclusion of this state in our analysis is necessary as, in Markov models, latent states should be mutually exclusive and exhaustive.

Table 2.2 presents the observed contract type distribution for the first month of the reference period according to the survey and the register data. The largest discrepancies occurs in the percentages of

individuals holding permanent and temporary contracts, and less in the "other" category. According to the survey data, in January 2007, 8% of the labour force was employed with a temporary contract, whereas in the register data this percentage is quite larger (12.3%).

Table 2.2
Distribution of contract types according to the survey and the register

	Survey	Register
Permanent	0.659	0.602
Temporary	0.080	0.123
Other	0.261	0.275
Total	1.0	1.0
Cases	3,887	11,632

Note: These frequency distributions refer to the first month of the reference period, January 2007. The LFS-sample is smaller than the PA-sample as only 3,887 LFS-respondents were interviewed for the first time in January 2007. The remaining respondents were interviewed in February and March 2007.

Table 2.3 cross-tabulates the contract type from the two sources for the pooled sample. This table confirms the large discrepancies between the two data sources reported by Statistics Netherlands. These discrepancies concern primarily individuals that are recorded as working on temporary contracts. More specifically, 50.2% of the individuals who are recorded as having a temporary contract in the register data appear to have a permanent contract in the survey. Smaller, but still existent, inconsistencies emerge for individuals that are recorded as having a permanent contract or as being in another state.

The inconsistencies in the classification of individuals that were presented in Table 2.3 have severe implications on the transitions between the different states. Table 2.4 presents the three-month transition rates for the cases with a valid observation from the LFS. This table indicates that the register data contain more transitions than the survey data. Specifically, from individuals that have a temporary contract in month $t-3$, 5.7% have a permanent contract in month t according to the survey data and 8.5% according to the register data.

Table 2.3
Cross-tabulation of contract type according to the survey and the register

Register data	Survey data			
	Permanent	Temporary	Other	Total
Permanent	0.944	0.039	0.017	1.0
Temporary	0.502	0.437	0.061	1.0
Other	0.081	0.030	0.889	1.0
Total	0.667	0.087	0.246	1.0
Cases	32,225	4,216	11,856	48,297

Note: The frequency distributions are calculated for the pooled sample. The grand total represents the number of LFS records included in our analysis in the pooled sample.

Table 2.4
Observed 3-month transitions in LFS and PA

Observed transitions from the survey data				
			Contract in t	
		Permanent	**Temporary**	**Other**
Contract in t-3	Permanent	0.981	0.009	0.010
	Temporary	0.057	0.889	0.054
	Other	0.017	0.035	0.948
	Total	0.674	0.089	0.237
Observed transitions from the register data				
			Contract in t	
		Permanent	**Temporary**	**Other**
Contract in t-3	Permanent	0.967	0.018	0.015
	Temporary	0.085	0.860	0.055
	Other	0.018	0.036	0.946
	Total	0.624	0.128	0.247

Note: For both tables, these are the transition rates over a 3-month period and for 34,820 cases of our pooled sample. These cases come from LFS-respondents that appear at least twice in our sample.

3 The hidden Markov model used to estimate the measurement error in the contract type

The model we use to estimate the error in the measurement of the contract type is a hidden or latent Markov model. This model has been used for the estimation of measurement error in variables from employment surveys (see, among others, van der Pol and Langeheine 1990; Rendtel et al. 1998; Bassi, Hagenaars, Croon and Vermunt 2000; Biemer and Bushery 2000; Biemer 2011; Pavlopoulos, Muffels and Vermunt 2012). Our application differs somewhat from these applications in that we have two measurements instead of a single one for the outcome variable; that is, the contract type from the PA and from the LFS. Other examples of applications of latent Markov models using multiple response variables are Langeheine (1994), Paas, Vermunt and Bijmolt (2007), Bartolucci, Lupparelli and Montanari (2009) and Manzoni, Vermunt, Luijkx and Muffels (2010).

Let C_{it} and E_{it} denote the observed state of person i at time point t according to the register and the survey, respectively, where $i = 1, \ldots, N$ and $t = 0, \ldots, T$. To deal with the fact that E_{it} is observed only every third month, we use the indicator variable δ_{it} which equals 1 if the survey information is available for the month concerned and 0 otherwise. In addition to the measurements from the register and survey, the hidden Markov model contains an unobserved variable representing an individuals' true contract type at time point t. We denote this latent state by X_{it}. Note that $C_{it}, E_{it},$ and X_{it} can take on three values representing the categories permanent, temporary, and other. We refer to a particular category of these variables by $c_t, e_t,$ and x_t, respectively.

The path diagram for the hidden Markov model of interest is depicted in Figure 3.1. For simplicity reasons, this path diagram refers only to individuals that entered the LFS-sample in a specific month. For this reason, from the four observations that are illustrated in the diagram, only those in months $t-3$ and t are non-missing for the LFS. As can be seen, the latent contract type X_{it} follows a first-order Markov process; that is, the true contract at time point $t, X_{it},$ is independent of the contract at time point $t', X_{it'},$ for $t' < t-1,$ conditionally on the state at $t-1, X_{i(t-1)}$. Another assumption is that the observed states

are independent of one another within and between time points, which is referred to as the local independence assumption or the assumption of independent classification errors (ICE). It can also be seen that E_{it} is observed only each third time point.

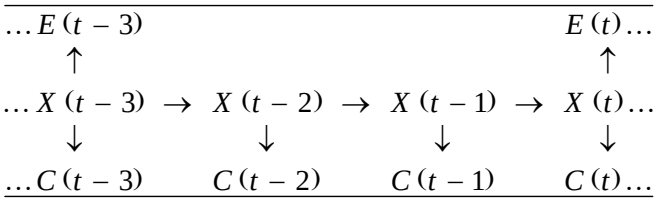


Figure 3.1 Path diagram for the hidden Markov model with two (partially) observed indicators

As indicated in the previous section, we use data for 15 months, which means that t runs from 0 to $T = 14$. The probability of following a certain observed path over the $T + 1$ months period can be expressed as follows:

$$P(\mathbf{C}_i = \mathbf{c}_i, \mathbf{E}_i = \mathbf{e}_i) = \sum_{x_0=1}^{3}\sum_{x_1=1}^{3}\cdots\sum_{x_T=1}^{3} P(X_{i0} = x_0)\prod_{t=1}^{T} P(X_{it} = x_t | X_{i(t-1)} = x_{t-1}) \prod_{t=0}^{T} P(C_{it} = c_t | X_{it} = x_t)\prod_{t=0}^{T} P(E_{it} = e_t | X_{it} = x_t)^{\delta_{it}}. \quad (3.1)$$

The relevant probabilities appearing in this equation are the initial state probabilities $P(X_{i0} = x_0)$, the time-specific transition probabilities $P(X_{it} = x_t | X_{i(t-1)} = x_{t-1})$, the measurement error probabilities for the register $P(C_{it} = c_t | X_{it} = x_t)$, and the measurement error probabilities for the survey $P(E_{it} = e_t | X_{it} = x_t)$.

So far, we assumed that the measurement error is uncorrelated across time points – that the ICE assumption holds – which may be unrealistic in our application. First of all, as indicated in the previous section, the measurement error in the register data is likely to be serially correlated; that is, when there is a mismatch between X_{it} and C_{it}, this increases the likelihood of having the same error at time point $t + 1$. This is the result of the fact that employers make mistakes in their registers which are not adapted until a regular control takes place. In the survey data and especially since we have prospective and not retrospective data, we have no reason to justify a similar "direct" autocorrelated error structure. However, the errors in the survey data may be correlated over time as a result of the fact that the probability of making an error may differ across groups of individuals, which is sometimes referred to as differential measurement error. Specifically, measurement error in the survey data is likely to be higher in sectors where mobility is common and ambiguity exists regarding the agreements between employers and workers, such as the health sector. Moreover, errors may be larger for young workers that care less about long-term employer relationships and therefore may have a less clear view than older respondents with respect to the formal arrangements they have on their contract. Figure 3.2 depicts the path diagram of the model correcting for possible heterogeneity and autocorrelation in the measurement error, where V represents the observed variables that introduce across-time correlation in the measurement error in the survey data.

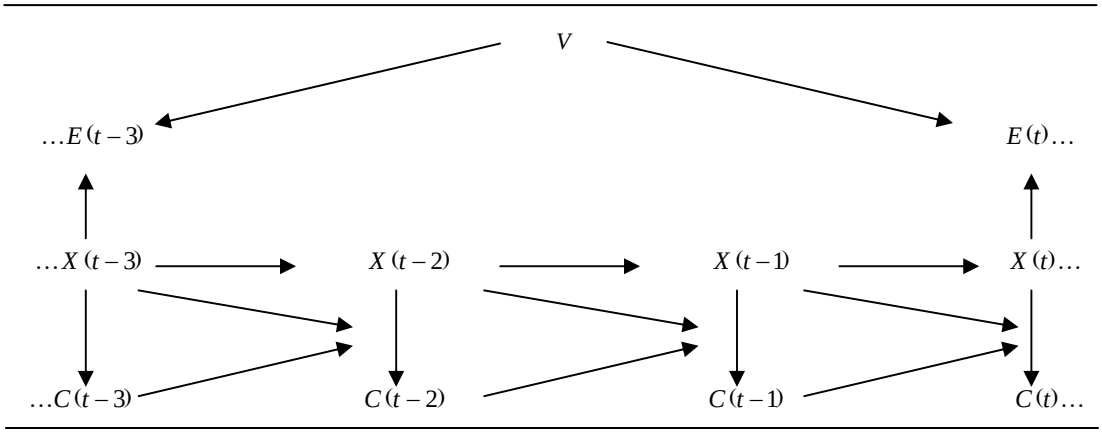


Figure 3.2 Path diagram for the hidden Markov model with two indicators and correlated errors

Because it is also important to control for the heterogeneity in the structural part of a Markov model (Shorrocks 1976), the model is further expanded with – possibly time-varying – observed variables affecting the initial state and latent transition probabilities, following the approach of Vermunt, Langeheine and Böckenholt (1999). We denote these control variables by $\mathbf{Z}_{it}$. However, these observed control variables cannot fully capture heterogeneity in the latent transition probabilities as these may be also affected by unobserved personal traits, such as motivation and ability. Following the most standard approach in the framework of hidden Markov models, we correct for unobserved heterogeneity by assuming that the population consists of a small number of latent classes with different initial state and transition probabilities (Poulsen 1990). In this way, we avoid the unattractive distributional assumptions on the latent variable that are adopted by continuous random-effects models (Heckman and Singer 1984; Vermunt 1997). The number of latent classes K can be determined using model fit indices.

In our mixed hidden Markov model, the joint probability of having a particular observed state path conditionally on predictor values can be expressed as:

$$
\begin{aligned}
P(\mathbf{C}_i = \mathbf{c}_i, \mathbf{E}_i = \mathbf{e}_i \mid \mathbf{V}_i, \mathbf{Z}_i) \;=\; & \sum_{k=1}^{K} \sum_{x_0=1}^{3} \sum_{x_1=1}^{3} \cdots \sum_{x_T=1}^{3} \pi_k P(X_{i0} = x_0 \mid \mathbf{Z}_{i0}, k) \\
& \prod_{t=1}^{T} P(X_{it} = x_t \mid X_{i(t-1)} = x_{t-1}, \mathbf{Z}_{it}, k) \\
& P(C_{i0} = c_0 \mid X_{i0} = x_0) \\
& \prod_{t=1}^{T} P(C_{it} = c_t \mid X_{it} = x_t, X_{i(t-1)} = x_{t-1}, C_{i(t-1)} = c_{t-1}) \\
& \prod_{t=0}^{T} P(E_{it} = e_t \mid X_{it} = x_t, \mathbf{V}_{it})^{\delta_{it}} ,
\end{aligned}
\tag{3.2}
$$

equation 3.2 specifies a finite mixture model with K latent classes to account for unobserved heterogeneity in the initial latent state and in the latent transition probabilities. π_k is the probability of belonging to the latent class k, $\mathbf{V}_{it}$ is the vector of covariates affecting the measurement error in the survey data (age and proxy interview) and $\mathbf{Z}_{it}$ is the vector of the covariates affecting the latent transition

probabilities (gender, age, education and country of origin). $\mathbf{Z}_{i0}$ is the vector of the values of these covariates in the initial time point.

Compared to equation 3.1, in equation 3.2, the error probabilities in the survey data are allowed to depend on covariates $(\mathbf{V}_{it})$. The covariate effects on these error probabilities are modelled using a logit model. Moreover, the error probabilities in the register data are allowed to depend on the lagged observed and lagged true contract type. Note that $X_{i(t-1)}$ and $C_{i(t-1)}$ can take on three values, which implies that there are nine (3 times 3) different sets of error probabilities in the register data, one for each possible combination of lagged observed and latent contract. Because it is not meaningful to estimate all these error probabilities freely, we used a more restricted model. More specifically, we define a logit model for $P(C_{it} = c_t | X_{it} = x_t, X_{i(t-1)} = x_{t-1}, C_{i(t-1)} = c_{t-1})$ of the form $\alpha_{c_t,x_t} + \beta_{c_t,c_{t-1},x_t,x_{t-1}}$, with $\beta_{c_t,c_{t-1},x_t,x_{t-1}}$ being a free parameter when $c_t = c_{t-1} \neq x_t = x_{t-1}$ (when the same error is made between adjacent time points) and otherwise being equal to 0. This model, which contains six additional parameters compared to a model without lagged effects on the misclassification probabilities, expresses that the likelihood of making a specific error depends on whether *the same error* was made at the previous time point. Similar restricted correlated error structures were used by Manzoni et al. (2010) in a latent Markov model for retrospectively collected responses.

The initial state and latent transition probabilities are also restricted using logit models, while for the latent transitions we use models with separate coefficients per origin state. The same set of covariates $(\mathbf{Z_{i0}}$ and $\mathbf{Z_{it}}$, respectively$)$ are introduced in the models estimating the initial state and latent transition probabilities. Note that the mixed hidden Markov model described in equation 3.2 assumes a first-order Markov process for the true states conditionally on the individuals' covariate values and time-constant unobserved effects, but this assumption does not need to hold after marginalizing over covariate values and latent classes. A simple first-order Markov model would be inappropriate for employment transitions especially at the month level. The reason is that there is duration dependence in unemployment. For example, it is unlikely to assume that an individual that was unemployed in months 3 to 9 has the same probability of being in a particular labour market state in month 10 as an individual that was unemployed only in month 9. However, in a hidden Markov model, the bias in the classification error due to the violation of the Markov assumption is minimal. Using simulations, Biemer and Bushery (2000) show that even in cases of a severe violation of the Markov assumption, in a hidden Markov model, the bias in the estimation of classification error in unemployment does not exceed 3%.

Maximum likelihood estimates of the model parameters are obtained using a variant of the Expectation-Maximization (EM) algorithm (Dempster, Laird and Rubin 1977) referred to as the forward-backward or Baum-Welch algorithm (Baum, Petrie, Soules and Weiss 1970). We use an extension of this algorithm for mixed latent Markov models with covariates as described – among others – in Vermunt, Tran and Magidson (2008) and Pavlopoulos et al. (2012). In the E-step, the expected complete data log-likelihood is computed, which involves computing the relevant marginal posterior probabilities for the latent classes and latent states. In the M-step, the model parameters are updated using standard algorithms for logistic regression analysis, where the marginal posterior probabilities are used as weights. This algorithm is implemented in the program Latent GOLD (Vermunt and Magidson 2008), which also provides standard errors for the model parameters (other popular programs for estimating latent Markov models are MPLUS, LEM and PANMARK).

Missing values due to the survey construction (as respondents are interviewed once per 3 months) are Missing Completely At Random (MCAR). Missing values due to attrition in the survey are treated as

Missing At Random (MAR). More specifically, following the standard manner within the ML estimation procedure, we maximize the log-likelihood for the incompletely observed data, which is obtained by integrating out the missing values. This procedure is valid under MAR.

As the LFS has a complex sampling design, the model has used the sampling weights of the survey, namely a single weight per observation. These weights are used in a pseudo ML estimation procedure, where the standard errors are adjusted for the weighting using a linearization estimator (Skinner, Holt and Smith 1989). Since these are trimester weights, they are not suitable for estimating population totals at the monthly level. However, as we use information from the register for all the LFS respondents that entered the survey in a certain trimester, these weights are appropriate for the estimation of hidden Markov models.

4 Results for the matched LFS and PA data

In total, we estimate the nine models that are presented in Table 4.1. All these models are first order hidden Markov models with two indicators for the contract type as presented in the Section 3. The error probabilities are time homogeneous. The (latent) transition probabilities are assumed to be time heterogeneous; that is, the transition logits are allowed to depend on time and time squared. These models are also finite mixture model that include three latent classes to control for unobserved heterogeneity in the initial latent state and in the latent transition probabilities. This number of latent classes was selected by comparing variants of Models B" and C with different number of latent classes (the results of these tests are available on request).

Models A', A" and A specify independent classification errors (ICEs) for the survey, the register and both datasets, respectively. Model B' specifies the error in the survey to depend on covariates V_{it} age and proxy interview, Model B" specifies serially correlated errors in the register, while Model B combines these two specifications. Models C' and C" extend Model B" by introducing predictors Z_{it} (gender, age, education and country of origin) for the transitions and for both the initial state and the transitions, respectively. Model C extends Model B by introducing the same predictors.

Table 4.1 presents the log-likelihood, the Bayesian Information Criterium (BIC), the Akaike Information Criterium (AIC) values and the number of parameters for nine of the models that were estimated with the matched LFS and PA data. In all models, the (latent) transition probabilities are assumed to be time heterogeneous; that is, the transition logits are allowed to depend on time and time squared.

Model A specifies that both the survey and the register data contain (independent) classification errors. As this model fits better than the restricted Models A' and A", which assume that only the survey (Model A') or only the register (Model A") contains errors, we conclude that there is evidence that both sources contain classification errors.

Models B', B" and B relax the ICE assumption for the survey, the register, and both the survey and the register, respectively. More specifically, the measurement error in the survey data is allowed to depend on the respondent's age and on whether the information was obtained using a proxy interview, and the measurement error in the register data is allowed to depend on the lagged latent and observed contract type. The latter is achieved by estimating a separate set of error probabilities for repeating *the same error*

across occasions. Restricted versions of Model B are estimated as well to examine whether the violation of the ICE assumption applies to the measurement error of only the survey data (Model B') or only the register data (Model B"). The fact that Model B" fits better than Models A and B' indicates that the ICE assumption should be relaxed for the indicator of the register data. Model B improves marginally the fit compared to Model B", which indicates that the ICE assumption for the survey indicator has also to be relaxed in a model without predictors for the transitions and for the initial state.

Table 4.1
Fit measures for eight models estimated with the matched LFS and PA data

Model	Log-likelihood	BIC (LL)	AIC (LL)	Parameters	L^2	df	P-value
A': ICE survey	-286,814	574,118	573,716	44	240,543.4	69,327	1.6e-18,454
A": ICE register	-454,196	908,882	908,480	44	575,307.7	69,327	8.5e-78,021
A: ICE both	-284,413	569,384	568,926	50	235,742.1	69,321	4.8e-17,717
B': A + non-ICE survey	-283,573	567,748	567,254	54	426,966.7	69,317	6.6e-50,302
B": A + non-ICE register	-246,054	492,732	492,220	56	435,025.8	69,315	2.9e-51,771
B: A + non-ICE both	-246,000	492,669	492,120	60	477,741.8	69,311	7.6e-59,639
C': B" + predictors transitions	-245,282	491,590	490,748	92	486,186.8	69,279	1.8e-61,222
C": B" + predictors initial & transitions	-241,990	485,140	484,189	104	479,603.4	69,267	4.9e-60,003
C: B + predictors initial & transitions	-242,006	485,217	484,229	108	479,635.2	69,263	1.2e-60,010

Note: Models A', A" and A specify independent classification errors (ICEs) for the survey, the register and both datasets, respectively. Model B' specifies the error in the survey to depend on age and proxy interview, Model B" specifies serially correlated errors in the register, while Model B combines these two specifications. Models C' and C" extend Model B" by introducing gender, age, education and country of origin as predictors for the transitions and for both the initial state and the transitions, respectively. Model C extends Model B by introducing the same predictors. All models are finite mixture models with 3 latent classes to correct for unobserved heterogeneity in the initial latent state and in the latent transition probabilities. Moreover, all models assume time heterogeneity for the latent transition probabilities. Specifically, we condition the latent transition probabilities on a linear trend for the month of the observation as well as on its square.

Finally, we extended Models B" and B by including covariates (gender, age, education and country of origin) in the models for the latent transition and the initial latent state probabilities (Model C" and C, respectively). Model C' is a restricted version of Model C" in which predictors are allowed to affect only the latent transition probabilities. The fact that Model C" fits better than Model B" and Model C' indicates that covariates have a significant effect on both the transitions and the initial states. The fact that, according to two of the three measures, Model C fits worse than Model C" means that the ICE assumption in the survey data should be retained in the model including predictors for the transitions and for the initial state (as the results of Model C show, the size of the measurement error in the survey data changes only marginally with age and proxy interview. This is further evidence in favor of retaining the ICE assumption for the survey indicator. Actually, the estimates for the size of the measurement error in both the survey and the register data and for the latent transition probabilities are very similar between the models C, C' and C". This shows that the results of our model are robust to small model misspecifications). In what follows, we present estimates derived from Model C" (the estimates from Models C and C' are available on request).

We investigated various alternative non-ICE models. Specifically, we studied whether the measurement error in the survey data differs for sectors with large contract and employment mobility, such as the health sector, but this did not turn out to be the case. For the register data, we looked at

alternative restricted specifications for the correlated errors, but these turned out to be worse in terms of model fit than the models from Table 4.1.

Now let us look at the amount of classification error in the two data sources. According to equation 3.2, for the survey and register data, this is represented by the probabilities $P(E_{it} = e_{it} | X_{it} = x_t)$ and $P(C_{it} = c_{it} | X_{it} = x_t, X_{i(t-1)} = x_{t-1}, C_{i(t-1)} = c_{t-1})$, respectively. The estimates from Model C" are presented in Tables 4.2 and 4.3. Specifically, Table 4.2 shows that permanent contracts and the other state are measured very accurately in LFS as almost all individuals are correctly classified. This is indicated by the large probabilities in the main diagonal of the table. Some error is found for individuals that have in reality a temporary contract. 12.5% of these individuals report that they have a temporary contract, while another 4.2% report being in another state.

Table 4.2
The size of the measurement error in the survey data according to Model C"

	Observed contract in t		
Latent contract in t	**Permanent**	**Temporary**	**Other**
Permanent	0.998	0.001	0.002
Temporary	0.125	0.832	0.042
Other	0.004	0.005	0.991

Note: Standard errors are always smaller than 0.0001.

Table 4.3 reports the estimated measurement-error probabilities for the register data, which according to equation 3.2 depend on the lagged observed and latent state. Due to the restrictions imposed (see Section 3), separate error (logit) parameters were estimated for repeating the same error between months $t-1$ and t. These situations correspond to the shaded cells in Table 4.3. As can be seen, the measurement errors are strongly autocorrelated; that is, if an error was made in month $t-1$ and if it was possible to repeat the same error (if one remained in the same latent state), the error almost surely persisted in month t. For instance, if an individual with a permanent contract in month $t-1$ was registered mistakenly as having a temporary contract and she had still a permanent contract in month t, then she had a 0.968 probability of being wrongly registered again as having a temporary contract in t. For the other five possible errors, the probability of a persisting measurement error is somewhat lower, but it is never below 0.84.

A different picture emerges when no error is made at time point $t-1$ or when an individual changes latent state between $t-1$ and t and therefore no error repetition is possible. In these cases, register data is almost error-free. For instance, when an individual was correctly registered as having a permanent contract in month $t-1$ and has a temporary contract at t, the contract type is registered correctly as temporary at t with a probability of 0.930. In practice, this means that the initial registration of the contract is crucial for the PA. If this registration is correct, then the registered contract type of the individual can be fully trusted until some true labour market change takes place. In contrast, if the contract type of the individual is initially registered wrongly, then this error will almost surely persist until the individual changes contract.

Table 4.3
Conditional probabilities of measurement error in register data according to Model C"

Observed contract in $t-1$	Latent contract in t	Latent contract in $t-1$	Observed contract in t		
			Permanent	Temporary	Other
Permanent	Permanent	Permanent	0.986	0.009	0.004
Permanent	Permanent	Temporary	0.986	0.009	0.004
Permanent	Permanent	Other	0.986	0.009	0.004
Permanent	Temporary	Permanent	0.045	0.930	0.025
Permanent	Temporary	Temporary	0.968	0.032	0.001
Permanent	Temporary	Other	0.045	0.930	0.025
Permanent	Other	Permanent	0.005	0.005	0.990
Permanent	Other	Temporary	0.005	0.005	0.990
Permanent	Other	Other	0.913	0.000	0.087
Temporary	Permanent	Permanent	0.027	0.973	0.000
Temporary	Permanent	Temporary	0.986	0.009	0.004
Temporary	Permanent	Other	0.986	0.009	0.004
Temporary	Temporary	Permanent	0.045	0.930	0.025
Temporary	Temporary	Temporary	0.045	0.930	0.025
Temporary	Temporary	Other	0.045	0.930	0.025
Temporary	Other	Permanent	0.005	0.005	0.990
Temporary	Other	Temporary	0.005	0.005	0.990
Temporary	Other	Other	0.001	0.842	0.157
Other	Permanent	Permanent	0.039	0.000	0.961
Other	Permanent	Temporary	0.986	0.009	0.004
Other	Permanent	Other	0.986	0.009	0.004
Other	Temporary	Permanent	0.045	0.930	0.025
Other	Temporary	Temporary	0.005	0.099	0.896
Other	Temporary	Other	0.045	0.930	0.025
Other	Other	Permanent	0.005	0.005	0.990
Other	Other	Temporary	0.005	0.005	0.990
Other	Other	Other	0.005	0.005	0.990

Note: Standard errors are always smaller than 0.0001.

To estimate the overall amount of error in the register data, we use the posterior probability of having a particular type of latent contract at each time point. This probability is estimated for all individuals in our sample by the hidden Markov model. These estimates are quite accurate as the classification error is only 0.016. The averages of these probabilities over individuals and time points are presented in Table 4.4. By comparing the probabilities in the main diagonal of Tables 4.1 and 4.4, we see that the error is larger in the register indicator than in the survey indicator. Specifically, individuals that are truly working on a temporary contract have a 0.237 probability of being registered as having a permanent contract (0.125 in the survey data) and a 0.079 probability of being registered as being in the other state in the PA (0.042 in the survey data). There is also some classification error for individuals that are truly working on a permanent contract, as they have a 0.081 probability of being registered as temporary workers and a 0.031 probability of being registered to another state.

Table 4.4
The size of the measurement error in the register data according to Model C"

Latent contract in t	Observed contract in t		
	Permanent	Temporary	Other
Permanent	0.888	0.081	0.031
Temporary	0.237	0.684	0.079
Other	0.032	0.017	0.951

Note: These probabilities are the average posterior probabilities of having a particular type of latent contract as estimated by Model C" with classification error 0.016.

We are not only interested in the measurement error itself, but also in how much it affects the estimate of the size of temporary employment. Using again the average posterior probabilities of having a particular type of latent contract, we estimate the size of temporary employment in the Netherlands. In Table 4.5, we compare the size of temporary employment as estimated by the hidden Markov model with the observed distributions of the contract type from the LFS and the PA. The average posterior probability of being in a temporary contract is 10.9% and lies in between the values obtained from LFS and PA.

Table 4.5
The average size of temporary employment according to Model C"

	Observed		Latent
	Survey	**Register**	
Permanent	0.667	0.597	0.634
Temporary	0.087	0.130	0.109
Other	0.246	0.273	0.257
Cases	48,297	174,480	174,480

Note: The latent probabilities are the average posterior probabilities of having a particular type of latent contract as estimated by Model C" with classification error 0.016.

Table 4.6 presents the evolution of the size of temporary employment according to the two data sources and according to the hidden Markov model. This table confirms the finding that the size of temporary employment according to our model is in between that of the register data and that of the survey data. It can also be seen that in the period of reference, the proportion of temporary employed increased. The small drop that is observed in the register data in January 2008 (month 13) compared to December 2007 (month 12) may be explained by the fact that many temporary contracts end on December 31st, and that, moreover, some of these contracts are converted into permanent contracts. The somewhat larger fluctuation in the size of temporary employment according to the survey data is due to the fact that respondents of the LFS are interviewed once per three months and thus the various monthly estimates come partly from different survey respondents.

Not only the aggregate change, but also the individual level change is important to be investigated; that is, the probability of making a transition from temporary to permanent employment and vice versa. These transition probabilities are presented in Table 4.7. More specifically, Table 4.7 presents the (average) latent transition probabilities obtained from Model C". The transition probabilities refer to a period of three months and are averaged over the 12 three-month periods in our data. If we compare the findings of Table 4.7 with those of Table 2.4, we see that the latent transitions probabilities are much smaller than those of both the register and the survey data. According to the latent transition probabilities, 3.2% of the individuals with a temporary contract were working with a permanent contract three months later, but according to the survey and register data, these percentages are 5.7% and 8.5%, respectively. This shows that measurement error inflates upwards the size of transition probabilities. Such an inflation would be clearly expected when errors are independent over time (Hagenaars 1990, 1994). When errors are not independent over time, as in our case, the expectation is less clear as errors may either increase or reduce the transitions, depending on the nature and the size of the association. The same pattern of underestimation of stability can be observed for the permanent contract state: 98.1% and 96.7% stayed in this state according to the survey and the register data, respectively, while the true stability is 98.7%.

Table 4.6
The evolution of the proportion of temporary employed for the period between January 2007 and March 2008

	Source		
Month	**Survey**	**Register**	**Latent**
1	0.080	0.123	0.102
2	0.082	0.124	0.103
3	0.085	0.123	0.102
4	0.084	0.128	0.103
5	0.084	0.129	0.103
6	0.090	0.129	0.104
7	0.089	0.130	0.105
8	0.087	0.131	0.106
9	0.091	0.135	0.110
10	0.087	0.134	0.112
11	0.088	0.135	0.114
12	0.091	0.135	0.114
13	0.090	0.131	0.116
14	0.089	0.131	0.118
15	0.096	0.132	0.121

Note: Survey data include trimonthly observations per individual, while register data include monthly observations per individual. The latent probabilities are the average posterior probabilities of having a particular type of latent contract as estimated by Model C" with classification error 0.016.

Table 4.7
Observed 3-months transitions in LFS and PA and latent transitions according to Model C

Latent transitions				
		Permanent	**Temporary**	**Other**
Contract in t-3	Permanent	0.987	0.006	0.007
	Temporary	0.032	0.931	0.037
	Other	0.009	0.030	0.961
	Total	0.634	0.110	0.256

Note: The latent probabilities are the average posterior probabilities of having a particular type of latent contract as estimated by Model C" with classification error 0.016.

5 Conclusions

In this paper, we investigated the measurement error in the type of the employment contract in the Dutch LFS by matching its longitudinal component from 2007 and early 2008 with a unique register dataset, the PA. We applied several hidden Markov models, in which the true contract type is treated as a latent state and in which the survey and register information serve as observed indicators of an individual's true contract. We modeled the measurement error in the two data sources by taking into account that the error in the register is correlated across occasions.

Our results show that the register data contain more error than the survey data, and therefore cannot be used as a golden standard. However, the improvement of the initial registration in the register data can significantly improve their quality as measurement error in the indicator of the contract type that comes from this dataset is serially correlated.

The measurement error results into an underestimation of the percentage of individuals that are working on a temporary contract. In the LFS this percentage is 8.9%, whereas after correction for measurement error this percentage rises to 10.9%. Another effect of measurement error is that it yields

severely overestimated transition probabilities. According to the LFS and PA, the transition probability between temporary to permanent employment in a three-month period is 5.7% and 8.5%, respectively, whereas the corresponding latent transition probability is only 3.2%. This finding is particularly important for Dutch policy makers as it clearly indicates that there is much less mobility from temporary to permanent employment than originally thought.

The results of this study remain fairly stable across the model specifications that we tested. This shows that the results are robust to small model misspecifications. However, results remain somehow dependant on model assumptions. Further sensitivity tests and applications can further verify the validity of our results. Future research may focus particularly on sensitivity tests with the use of Monte Carlo simulations.

Acknowledgements

The authors are thankful to Statistics Netherlands for providing access to the data of this article. The authors are also thankful to Frank van der Pol, Wendy Smits, Ruben van Gaalen and to the participants of the ESPE and EALE conferences as well as to the participants of the SILC research group of the VU University Amsterdam for the useful comments and suggestions. The contribution of Jeroen Vermunt was supported by the Netherlands Organization for Scientific Research (NWO) [VICI grant number 453-10-002].

References

Abowd, J.M., and Stinson, M.H. (2005). Estimating measurement error in SIPP annual job earnings: A comparison of census survey and SSA administrative data. Technical Paper, *U.S. Census Bureau*.

Bartolucci, F., Lupparelli, M. and Montanari, G.E. (2009). Latent markov model for longitudinal binary data: An application to the performance evaluation of nursing homes. *Annals of Applied Statistiscs*, 3(2), 611-636.

Bassi, F., Hagenaars, J.A., Croon, M.A. and Vermunt, J.K. (2000). Estimating true changes when categorical panel data are affected by uncorrelated and correlated classification errors. *Sociological Methods and Research*, 29(2), 230-268.

Baum, L.E., Petrie, T., Soules, G. and Weiss, N. (1970). A maximization technique occuring in the statistical analysis of probabilistic functions of markov chains. *The Annals of Mathematical Statistics*, 41(1), 164-171.

Bentolila, S., and Bertola, G. (1990). Firing costs and labour demand: How bad is eurosclerosis? *The Review of Economic Studies*, 57(3), 381-402.

Biemer, P. (2004). An analysis of classification error for the revised Current Population Survey employment questions. *Survey Methodology*, 30, 2, 127-140.

Biemer, P. (2011). *Latent Class Analysis of Survey Error*. New Jersey: John Wiley & Sons, Inc.

Biemer, P.P., and Bushery, J.M. (2000). On the validity of markov latent class analysis for estimating classification error in labor force data. *Survey Methodology*, 26, 2, 139-152.

Bollinger, C.R. (1996). Bounding mean regressions when a binary regressor is mismeasured. *Journal of Econometrics*, 73(2), 387-399.

Booth, A.L. (1997). An analysis of firing costs and their implications for unemployment policy. In *Unemployment Policy*, (Eds., D.J. Snower and G. de la Dehesa). Cambridge: Cambridge University Press.

Bound, J., Brown, C., Duncan, G.J. and Rodgers, W.L. (1994). Evidence on the validity of cross-sectional and longitudinal labor market data. *Journal of Labor Economics*, 12(3), 345-368.

Bound, J., Brown, C. and Mathiowetz, N. (2001). Measurement error in survey data. In *Handbook of econometrics*, (Eds., J.J. Heckman and E. Leamer), Amsterdam: Elsevier, 5, 3705-3843.

Brown, C., and Medoff, J.L. (1996). Employer characteristics and work environment. *Annales D'Économie et de Statistique*, 41, 275-298.

Cahuc, P., and Postel-Vinay, F. (2002). Temporary jobs, employment protection and labor market performance. *Labour Economics*, 9(1), 63-91.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1), 1-38.

Duncan, G.J., and Hill, D.H. (1985). An investigation of the extent and consequences of measurement error in labor-economic survey data. *Journal of Labor Economics*, 3(3), 508-522.

Gottschalk, P. (2005). Downward nominal-wage flexibility: Real or measurement error. *Review of Economics and Statistics*, 87(3), 556-568.

Gottschalk, P., and Huynh, M. (2010). Are earlings inequality and mobility overstated? The impact of non-classical measurement error. *Review of Economics and Statistics*, 92(2), 302-315.

Hagenaars, J.A. (1990). Categorical Longitudinal Data Log-Linear Panel, Trend and Cohort Analysis. Newbury Park, CA: Sage Publications.

Hagenaars, J.A. (1994). Latent variables in log-linear models of repeated observations. In *Latent Variable Analysis: Applications for Developmental Research*, (Eds., A. von Eye and C.C. Clogg). Thousand Oaks, CA: Sage Publications, 329-352.

Heckman, J.J., and Singer, B.L. (1984). A method for minimising the impact of distributional assumptions in econometric models for duration data. *Econometrica*, 52 (2), 271-320.

Hilbers, P., Houwing, H. and Kösters, L. (2011). De flexibele schil – overeenkomsten en verschillen tussen CBS- en UWV-cijfers [the flexible periphery – similarities and differences between CBS and UWV-data]. In *Sociaaleconomische Trends, 2e Kwartaal 2011 [Socioeconomic Trends, 2nd Trimester 2011]*, (Eds., B. Hermans et al.). Den Haag/Heerlen: Statistics Netherlands, 26-33.

Kapteyn, A., and Ypma, J.Y. (2007). Measurement error and misclassification: A comparison of survey and register data. *Journal of Labor Economics*, 25(3), 513-551.

Langeheine, R. (1994). Latent variable markov models. In *Latent Variables Analysis. Applications for Developmental Research*, (Eds., A. von Eye and C. Clogg). Thousand Oaks, California: Sage Publications, 373-395.

Manzoni, A., Vermunt, J.K., Luijkx, R. and Muffels, R. (2010). Memory bias in retrospectively collected employment careers: A model-based approach to correct for measurement error. *Sociological Methodology*, 40(1), 39-73.

Mars, G. (2011, December). Cijfers over Flexibele Arbeidsrelaties - Confrontatie Van Bronnen en Defiities [Figures on Flexible Labour Relations - Confrontation of Sources and Definitions]. Statistics Netherlands, report nr SAH-2011-H11. The Hague/Heerlen.

Mathiowetz, N.A. (1992). Errors in reports of occupations. *Public Opinion Quarterly*, 56(3), 352-355.

OECD (2002). *Employment Outlook* 2002. Paris: Author.

OECD (2012). *Country Statistical Profiles*. OECD Database: retrieved on 2012/12/16 from http://stats.oecd.org/.

Paas, L.J., Vermunt, J.K. and Bijmolt, T.H. (2007). Discrete-time discrete-state latent markov modelling for assessing and predicting household acquisitions of financial products. *Journal of the Royal Statistical Society, Series A*, 170(4), 955-974.

Pavlopoulos, D., Muffels, R. and Vermunt, J.K. (2012). How real is mobility between low pay, high pay and non-employment. *Journal of Royal Statistical Society, Series A*, 175(3), 749-773.

Pischke, J.-S. (1995). Measurement error and earnings dynamics: Some estimates from the PSID validation study. *Journal of Business and Economic Statistics*, 13(3), 305-314.

Poulsen, C.S. (1990). Mixed markov and latent markov modelling applied tobrand choice behaviour. *International Journal of Research in Marketing*, 7(1), 5-19.

Rendtel, U., Langeheine, R. and Berntsen, R. (1998). The estimation of poverty dynamics using different measurements of household income. *Review of Income and Wealth*, 44(1), 81-98.

Rodgers, W.L., Brown, C. and Duncan, G.J. (1993). Errors in survey reports of earnings, hours worked, and hourly wages. *Journal of the American Statistical Association*, 88(3), 345-368.

Sels, L., and Van Hootegem, G. (2001). Seeking the balance between flexibility and security: A rising issue in the low countries. *Work, Employment and Society*, 15(2), 327-352.

Shorrocks, A.F. (1976). Income mobility and the markov assumption. *Economic Journal*, 86, 566-578.

Skinner, C.J., Holt, D. and Smith, T.M.F. (1989). *Analysis of Complex Surveys*. Wiley.

van der Pol, F., and Langeheine, R. (1990). Mixed markov latent class models. *Sociological Methodology*, 20, 213-247.

Vermunt, J.K. (1997). *Log-Linear Models for Event Histories*. London: SAGE publications.

Vermunt, J.K., and Magidson, J. (2008). *LG - Syntax User's Guide: Manual for Latent GOLD 4.5 Syntax Module*. Belmont Massachussetts: Statistical Innovations Inc.

Vermunt, J.K., Langeheine, R. and Böckenholt, U. (1999). Discrete-time discrete-state latent markov models with time-constant and time-varying covariates. *Journal of Educational and Behavioral Statistics*, 24, 178-205.

Vermunt, J.K., Tran, B. and Magidson, J. (2008). Latent class models in longitudinal research. In *Handbook of Longitudinal Research: Design, Measurement, and Analysis*, (Ed., S. Menard). Burlington, MA: Elsevier, 373-385.

Survey Methodology, June 2015
Vol. 41, No. 1, pp. 215-236
Statistics Canada, Catalogue No. 12-001-X

Generalized framework for defining the optimal inclusion probabilities of one-stage sampling designs for multivariate and multi-domain surveys

Piero Demetrio Falorsi and Paolo Righi[1]

Abstract

This paper introduces a general framework for deriving the optimal inclusion probabilities for a variety of survey contexts in which disseminating survey estimates of pre-established accuracy for a multiplicity of both variables and domains of interest is required. The framework can define either standard stratified or incomplete stratified sampling designs. The optimal inclusion probabilities are obtained by minimizing costs through an algorithm that guarantees the bounding of sampling errors at the domains level, assuming that the domain membership variables are available in the sampling frame. The target variables are unknown, but can be predicted with suitable super-population models. The algorithm takes properly into account this model uncertainty. Some experiments based on real data show the empirical properties of the algorithm.

Key Words: Optimal Allocation; Multi-way stratification; Domain estimates; Balanced Sampling.

1 Introduction

Surveys conducted in the context official statistics commonly produce a large number of estimates relating to both different parameters of interest and highly detailed estimation domains. When the domain indicator variables are available for each sampling unit in the sampling frame, the survey sampling designer could attempt to select a sample in which the size for each domain is fixed. Thus, direct estimates can be obtained for each domain and sampling errors at the domain level would be controlled. We hereby present a *unified* and *general* framework for defining the *optimal inclusion probabilities* for *uni-stage sampling designs* when the domain membership variables are known at the design stage. This case may be the most recurrent scenario in establishment surveys and other survey contexts, such as agricultural surveys or social surveys if the domains are geographical (e.g., type of municipality, region, province, etc.). The growing development of data integration among administrative registers and survey frames may also increase the applicability of the approach presented herein in social surveys too. The proposal may be useful for planning an optimal second phase survey if, during the first phase, the domain membership variables have been collected.

The problem of defining optimal sampling designs has been addressed in some recent papers. Gonzalez and Eltinge (2010) present an interesting overview of the approaches for defining optimal sampling strategies. The optimization problem is usually dealt with in stratified sampling designs with a fixed sample size in each stratum. The optimal allocation in stratified samplings for a univariate population is well-known in sampling literature (Cochran 1977). In multivariate cases, where more than one characteristic is to be measured on each sampled unit, the optimal allocation for individual characteristics is of little practical use unless the various characteristics under study are highly correlated. This is because an allocation which is optimal for one characteristic is generally far from being optimal for others. The

1. Piero Demetrio Falorsi, FAO, Viale delle Terme di Caracalla, Roma. E-mail: piero.falorsi@fao.org; Paolo Righi, ISTAT Via C. Balbo 16, 00184 Roma. E-mail: parighi@istat.it.

multidimensionality of the problem leads to definition of a compromise allocation method (Khan, Mati and Ahsan 2010) with a loss of precision compared to the individual optimal allocations. Several authors have discussed various criteria for obtaining a feasible compromise allocation - see e.g., Kokan and Khan (1967), Chromy (1987), Bethel (1989), Falorsi and Righi (2008), Falorsi, Orsini and Righi (2006) and Choudhry, Rao and Hidiroglou (2012).

Recently, some papers have focused on finding optimal inclusion probabilities in balanced sampling (Tillé and Favre 2005; Chauvet, Bonnéry and Deville 2011), a general class of sampling designs that includes stratified sampling designs as special cases. In particular, Chauvet et al. (2011) propose the adoption of the fixed point algorithm for defining the optimal inclusion probabilities. Nevertheless, the above mentioned papers do not address the case in which the balancing variables depend on the inclusion probabilities and present only a partial solution to the problem related to the fact that the sampling variance is an **implicit** function of the inclusion probabilities. Choudhry et al. (2012) propose an optimal allocation algorithm for domain estimates in stratified sampling (if the estimation domains do not cut across the strata). Their algorithm represents a special case of the approach proposed herein. The methodological setting illustrated here is a substantial improvement with respect to the earlier version of the methodology described in Falorsi and Righi (2008) which only accounted for the case in which the values of the variables of interest were known and the measure of accuracy was expressed by the design variance; furthermore, the previous version did not consider the fact that the design variance, bounded in the optimization problem, is an implicit function of the inclusion probabilities. This paper studies the more realistic case in which the variables of interest are not known and must be estimated. Moreover, it explicitly deals with the problem that the anticipated variances are implicit functions of the inclusion probabilities. The new optimization algorithm can be easily performed because it is based on a general decomposition of the measure of accuracy. A general sampling design which includes most of the one-stage sampling designs adopted in actual surveys is proposed, e.g., Simple Random Sampling Without Replacement (SRSWOR), Stratified SRSWOR, Stratified PPS, Designs with incomplete stratification, etc. The framework is based on a joint use of *balanced sampling designs* (Deville and Tillé 2004) which, depending upon the different definitions of the balancing equations, represents a wide-ranging sampling design and *superpopulation models for predicting* the unknown values of the variables of interest. The paper is structured as follows. Section 2 introduces definitions and notations. Section 3 and Section 4 illustrate the sampling design and the Anticipated Variance. The algorithm for defining the optimal inclusion probabilities is described in Section 5. In Section 6, some experiments based on real business data show the empirical properties of the algorithm. The conclusions are given in Section 7.

2 Definitions and notation

In this section, we introduce the concepts of ***estimation domain*** and ***planned domain*** which play a key role in the framework presented herein.

Let U be the reference population of N elements and let $U_d\ (d = 1,\ldots,D)$ be an ***estimation domain***, i.e., a generic sub-population of U with N_d elements, for which separate estimates must be calculated. Let y_{rk} denote the value of the r^{th} $(r = 1,\ldots,R)$ variable of interest attached to the k^{th} population unit and let γ_{dk} denote the domain membership indicator for unit k defined as

$$\gamma_{dk} = \begin{cases} 1 & \text{if } k \in U_d \\ 0 & \text{otherwise} \end{cases}. \tag{2.1}$$

We assume that the γ_{dk} values are available in the sampling frame and more than one value γ_{dk} $(d = 1,\ldots,D)$ can be 1 for each unit k; therefore, the estimation domains can overlap.

The parameters of interest are the $D \times R$ domain totals

$$t_{(dr)} = \sum_{k \in U} y_{rk}\gamma_{dk}\ (r = 1,\ldots,R; d = 1,\ldots,D). \tag{2.2}$$

Let $p(\cdot)$ be a single-stage without replacement sampling design and $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_k,\ldots,\pi_N)'$ be the N-vector of inclusion probabilities. Let s be the sample selected with probability $p(s)$. Denote by U_h $(h = 1,\ldots,H)$ the subpopulation of size $N_h = \sum_{k \in U_h} \delta_{hk}$ where $\delta_{hk} = 1$ if $k \in U_h$ and $\delta_{hk} = 0$ otherwise.

We focus on fixed size sampling designs which are those satisfying

$$\sum_{k \in s} \boldsymbol{\delta}_k = \mathbf{n}, \tag{2.3}$$

where $\boldsymbol{\delta}_k = (\delta_{1k},\ldots,\delta_{hk},\ldots,\delta_{Hk})'$ and $\mathbf{n} = (n_1,\ldots,n_h,\ldots,n_H)'$ is the vector of integer numbers defining the sample sizes fixed at the design stage. Since the sample size n_h, corresponding to U_h, does not vary among sample selections, the subpopulation U_h will be referred to as a ***planned domain*** in the sequel. A necessary but not sufficient condition for ensuring that (2.3) is satisfied is that the vector $\boldsymbol{\pi}$ is such that

$$\sum_{k \in U} \pi_k \boldsymbol{\delta}_k = \mathbf{n}. \tag{2.4}$$

In our setting, the planned domains can overlap; therefore, the unit k may have more than one value $\delta_{hk} = 1$ (for $h = 1,\ldots,H$). Let us suppose that the δ_{hk} values are known, and available in the sampling frame, for all population units. We suppose furthermore that the $N \times H$ matrix $(\boldsymbol{\delta}_1,\ldots,\boldsymbol{\delta}_k,\ldots,\boldsymbol{\delta}_N)'$ is non-singular.

The planned domains and their relationship with the estimation domains play a central role in our generalized framework. We assume that the estimation domains may be defined as an aggregation of complete planned domains, which ensure that the *expected* sample size in the d^{th} estimation domain U_d, say n_d, can be obtained as a simple aggregation of the expected sample sizes of the planned domains that are included within it. Finally, let $\hat{t}_{(dr)}$ be the Horvitz-Thompson (HT) estimator of $t_{(dr)}$ with

$$\hat{t}_{(dr)} = \sum_{k \in s} \frac{1}{\pi_k} y_{rk}\gamma_{dk}. \tag{2.5}$$

An example from business surveys. Suppose that the survey estimates must be calculated separately considering three domain types: *region* (with 20 modalities), *economic activity* (2 modalities: goods and services) and *enterprise size* (3 modalities: small, medium and large enterprises). That is, there are

$D = 20 + 2 + 3 = 25$ possible overlapping estimation domains. The planned domains can be defined with different options.

Option 1. The single planned domain U_h is identified by a specific intersection of the categories of the estimation domains. In this case $H = 20 \times 2 \times 3 = 120$ planned domains are defined. They represent a specific partition of U. The planned domains do not overlap and $\sum_h \delta_{hk} = 1$.

Option 2. The planned domains U_h coincide with the estimation domains. Therefore, $H = D = 25$ and the $\boldsymbol{\delta}'_k$ are defined as vectors with three 1's, so that $\sum_h \delta_{hk} = 3$. Recall that the planned domains overlap.

Option 3. The planned domains U_h are defined as (*i*) region by economic activity and (*ii*) economic activity by enterprise size; then, $H = (20 \times 2) + (2 \times 3) = 46$ with $\sum_h \delta_{hk} = 2$.

Other intermediate relationships among estimation and planned domains are possible.

It is emphasised that the planned domains represent the basis for defining broad classes of sampling designs. For instance, ***stratified sampling designs*** require that the planned domains do not overlap, as $\sum_h \delta_{hk} = 1$ and each U_h is referred to as a stratum. Therefore, Option 1 in the example above leads us to define a stratified sampling design. Furthermore, the strata defined as in Option 1 are the basis of the so-called "multi-way stratified sampling design" (Winkler 2001).

If $\sum_h \delta_{hk} > 1$, the sample sizes of the planned domains identified in Option 1 (strata) are not strictly controlled. Nevertheless, the sample sizes are still controlled at an aggregated level. In Option 2 of the example above, the sample sizes are controlled only for the estimation domains; while in Option 3, the sample sizes are controlled for the subsets of two different partitions, defined by (*i*) the region by economic activity and (*ii*) the economic activity by enterprise size. On the basis of the Winkler's definition, we denote the designs using these types of planned domains as ***Incomplete multi-way Stratified Sampling*** **(ISS)** ***designs***.

3 Sampling

Let $\mathbf{z}_k$ be a vector of auxiliary variables available for all $k \in U$. A sampling design $p(s)$ is said to be balanced on the auxiliary variables if and only if it satisfies the following ***balancing equations***

$$\sum_{k \in s} \frac{\mathbf{z}_k}{\pi_k} = \sum_{k \in U} \mathbf{z}_k \tag{3.1}$$

for each sample s such that $p(s) > 0$ (Deville and Tillé 2004). Depending on the auxiliary variables and the inclusion probabilities, equation (3.1) can be exactly or approximately satisfied in each possible sample; therefore, a balanced sampling design does not always exist. By specifying

$$\mathbf{z}_k = \pi_k \boldsymbol{\delta}_k, \tag{3.2}$$

equations (3.1) become

$$\sum_{k\in s}\boldsymbol{\delta}_k = \sum_{k\in U}\pi_k\boldsymbol{\delta}_k. \tag{3.3}$$

In this case, the balancing equations state that the sample size achieved in each subpopulation U_h is equal to the expected size. In different contexts, Ernst (1989) and Deville and Tillé (2004; page 905 Section 7.3), have proved that, (*i*) with the specification (3.2) and (*ii*) if the vector of the expected sample sizes, given by $\mathbf{n} = \sum_{k\in U}\pi_k\boldsymbol{\delta}_k$, includes only integer numbers, then a balanced sampling design always exists. Specification (3.2) defines sampling designs that guarantee equation (2.4), upon which we wish to focus on. Deville and Tillé (2004, pages 895 and 905), Deville and Tillé (2005, page 577) and Tillé (2006, page 168) have shown that several customary sampling designs may be considered as special cases of balanced sampling, by properly defining the vectors $\boldsymbol{\pi}$ and $\boldsymbol{\delta}_k$ of equation (3.2). These issues are illustrated in Remark 4.2 and in Section 6. Balanced samples may be drawn by means of the Cube method (Deville and Tillé 2004). This strongly facilitates the sample selection of incomplete stratified sampling designs that overcome the computational drawbacks of methods based on linear programming algorithms (Lu and Sitter 2002). The Cube method satisfies (3.1) exactly when (3.2) holds and $\mathbf{n}$ is a vector of integers. In the cases of SRSWOR and SSRSWOR, the standard sample selection methods can be used, as well as the Cube method. Deville and Tillé (2005) propose as approximation of the variance for the HT estimator, in the balanced sampling

$$E_p\left(\hat{t}_{(dr)} - t_{(dr)}\right)^2 \cong [N/(N-H)]\left[\sum_{k\in U}(1/\pi_k - 1)\,\eta^2_{(dr)k}\right] \tag{3.4}$$

where E_p denotes the sampling expectation and

$$\eta_{(dr)k} = y_{rk}\gamma_{dk} - \pi_k\boldsymbol{\delta}_k'\left[\mathbf{A}(\boldsymbol{\pi})\right]^{-1}\sum_{j\in U}\pi_j\left(1/\pi_j - 1\right)\boldsymbol{\delta}_j y_{rk}\gamma_{dk} \tag{3.5}$$

with

$$\mathbf{A}(\boldsymbol{\pi}) = \sum_{j\in U}\boldsymbol{\delta}_j\boldsymbol{\delta}_j'\pi_j\left(1-\pi_j\right). \tag{3.6}$$

Recently, the simulation results in Breidt and Chauvet (2011) confirm that equation (3.4) represents a good approximation of the sampling variance when the balanced equations are satisfied exactly. Variance estimation is studied in Deville and Tillé (2005).

4 Anticipated variance

Prior to sampling, the y_{rk} values are not known and the variance expressed in formula (3.4) cannot be used for planning the sampling precision at the design phase. In practice, it is necessary to either obtain some proxy values or predict the y_{rk} values based on superpopulation models that exploit auxiliary information. The increasing availability of auxiliary information (deriving by integration of administrative registers and survey frames) facilitates the use of predictions. Under a model-based inference, the y_{rk} values are assumed to be the realization of a superpopulation model M. The model we study has the following form:

$$\begin{cases} y_{rk} = f_r(\mathbf{x}_k;\boldsymbol{\beta}_r) + u_{rk} \\ E_M(u_{rk}) = 0 \ \ \forall k;\ E_M(u_{rk}^2) = \sigma_{rk}^2;\ E_M(u_{rk},u_{rl}) = 0 \ \ \forall k \neq l \end{cases}, \tag{4.1}$$

where $\mathbf{x}_k$ is a vector of predictors (available in the sampling frame), $\boldsymbol{\beta}_r$ is a vector of regression coefficients and $f_r(\mathbf{x}_k;\boldsymbol{\beta}_r)$ is a known function, u_{rk} is the error term and $E_M(\cdot)$ denotes the expectation under the model. The parameters $\boldsymbol{\beta}_r$ and the variances σ_{rk}^2 are assumed to be known, although in practice they are usually estimated. The model (4.1) is variable-specific and different models for different variables may be used and this does not create additional difficulty. As a measure of uncertainty, we consider the *Anticipated Variance* (AV) (Isaki and Fuller 1982):

$$\mathrm{AV}(\hat{t}_{(dr)}) = E_M E_p(\hat{t}_{(dr)} - t_{(dr)})^2. \tag{4.2}$$

A general expression for the AV under linear models was derived by Nedyalkova and Tillé (2008). Their formulation is obtained by considering a linear function $f_r(\cdot)$ and a unique set of auxiliary variables, $\mathbf{x}_k$, used for both the prediction of the y values and for balancing the sample. In our context, we have introduced $\mathbf{x}_k$ and $\mathbf{z}_k = \pi_k\boldsymbol{\delta}_k$, highlighting that the auxiliary variables can be different for prediction and balancing. The variables $\mathbf{x}_k$ must be as predictive of y_{rk} as possible, while the variables $\mathbf{z}_k$ play an instrumental role in controlling the sample sizes for sub-populations.

In the context considered here, inserting the approximate variance (3.4) in the equation (4.2), we obtain the approximate expression of the AV:

$$\mathrm{AAV}(\hat{t}_{(dr)}) = [N/(N-H)]\sum_{k\in U}(1/\pi_k - 1)E_M(\eta_{(dr)k}^2), \tag{4.3}$$

where the terms $\eta_{(dr)k}^2$ in (3.4) are replaced by $E_M(\eta_{(dr)k}^2)$. By defining

$$\tilde{y}_{rk} = f_r(\mathbf{x}_k;\boldsymbol{\beta}_r), \tag{4.4}$$

the equation (4.3) may be reformulated as

$$\mathrm{AAV}(\hat{t}_{(dr)}) = [N/(N-H)]\left[\sum_{k\in U}\frac{1}{\pi_k}(\tilde{y}_{rk}^2 + \sigma_{rk}^2)\gamma_{dk} - \sum_{k\in U}(\tilde{y}_{rk}^2 + \sigma_{rk}^2)\gamma_{dk} - \mathrm{AAV}_{3(dr)}\right], \tag{4.5}$$

where the third variance component of $\mathrm{AAV}(\hat{t}_{(dr)})$ is

$$\begin{aligned} \mathrm{AAV}_{3(dr)} &= \sum_{k\in U}(1-\pi_k)a_{(dr)k}(\boldsymbol{\pi})[2\tilde{y}_{rk}\gamma_{dk} - \pi_k a_{(dr)k}(\boldsymbol{\pi})] \\ &+ \sum_{k\in U}(1-\pi_k)[2b_{(dr)k}(\boldsymbol{\pi}) - \pi_k c_{(dr)k}(\boldsymbol{\pi})] \end{aligned} \tag{4.6}$$

and $a_{(dr)k}(\boldsymbol{\pi})$, $b_{(dr)k}(\boldsymbol{\pi})$ and $c_{(dr)k}(\boldsymbol{\pi})$ are real numbers defined respectively in equations (A1.4), (A1.7) and (A1.8) of Appendix A1.

Remark 4.1. Expression (4.5) is a cumbersome formula but, for all practical purposes, calculations may be simplified by considering a slight upward approximation by setting $b_{(dr)k}(\boldsymbol{\pi}) = c_{(dr)k}(\boldsymbol{\pi}) = 0$ in (4.6). The proof is given in Appendix A3. An upward approximation is a safe choice in this setting, since it averts from the risk of defining an insufficient sample size for the expected accuracy.

Remark 4.2. The SSRSWOR design is obtained if the planned domains define a unique partition of population (Option 1 of the example in Section 2) and the model (4.1) is specified so that the predicted values are: $\tilde{y}_{rk} = \bar{Y}_{rh}$ with $\sigma^2_{rk} = \sigma^2_{rh}$ (for $k \in U_h$). The AAV becomes

$$\mathrm{AAV}(\hat{t}_{(dr)}) = [N/(N-H)] \sum_{d=1}^{D} \sum_{h \in H_d} \sigma^2_{rh} N_h (N_h/n_h - 1), \tag{4.7}$$

where H_d is the set of planned domains included in U_d (see Appendix A4). Note that the expression (4.7) agrees with the *Result 2* of Nedyalkova and Tillé (2008), but for the term $N/(N-H)$. If $[N/(N-H)](1/N_h) \approx 1/(N_h - 1)$ the expression (4.7) would approximate the variance of the HT estimate in the SSRSWOR design. The above approximation is proved true when the number of domains H remains small compared to the overall population size N, and when the domain sizes N_h are large.

5 Determination of the optimal inclusion probabilities

The vector of π-values is determined by solving the following optimization problem:

$$\begin{cases} \mathrm{Min}\left(\sum_{k \in U} \pi_k c_k\right) \\ \mathrm{AAV}(\hat{t}_{(dr)}) \le \bar{V}_{(dr)} \quad (d = 1, \ldots, D; r = 1, \ldots, R), \\ 0 < \pi_k \le 1 \qquad\qquad (k = 1, \ldots, N) \end{cases} \tag{5.1}$$

where c_k is the cost for collecting information from unit k and $\bar{V}_{(dr)}$ is a fixed variance threshold corresponding to $\hat{t}_{(dr)}$. System (5.1) minimizes the expected cost ensuring that the anticipated variances are bounded and that the inclusion probabilities lie between 0 and 1. If all the c_k values are constants equal to 1, then the problem (5.1) minimizes the sample size. We note that in problem (5.1) the variances σ^2_{rk} in $\mathrm{AAV}(\hat{t}_{(dr)})$ are treated as known; in practice they must be estimated. In Section 6, an empirical evaluation is conducted in order to study the sensitivity of the overall sample size with different estimated values of σ^2_{rk}.

To solve (5.1), we rearrange the inequality constraints to obtain

$$\sum_{k \in U} \frac{(\tilde{y}^2_{rk} + \sigma^2_{rk})\gamma_{dk}}{\pi_k} \le \frac{N-H}{N} \bar{V}_{(dr)} + \sum_{k \in U} (\tilde{y}^2_{rk} + \sigma^2_{rk})\gamma_{dk} + \mathrm{AAV}_{3(dr)}. \tag{5.2}$$

By fixing the values of $\mathrm{AAV}_{3(dr)}$ appropriately, the optimization problem becomes a classical Linear Convex Separate Problem (LCSP; Boyd and Vanderberg 2004). Figure 5.1 depicts the flow chart of the algorithm (A prototype software implementing the algorithm is available at http://www.istat.it/it/strumenti/metodi-e-software/software.), which is organized into two nested loops: the **Outer Loop** (OL) and the **Inner Loop** (IL). The two loops are updated according to a *fixed point* algorithm scheme. The convergence under some approximations is shown in Appendix A2.

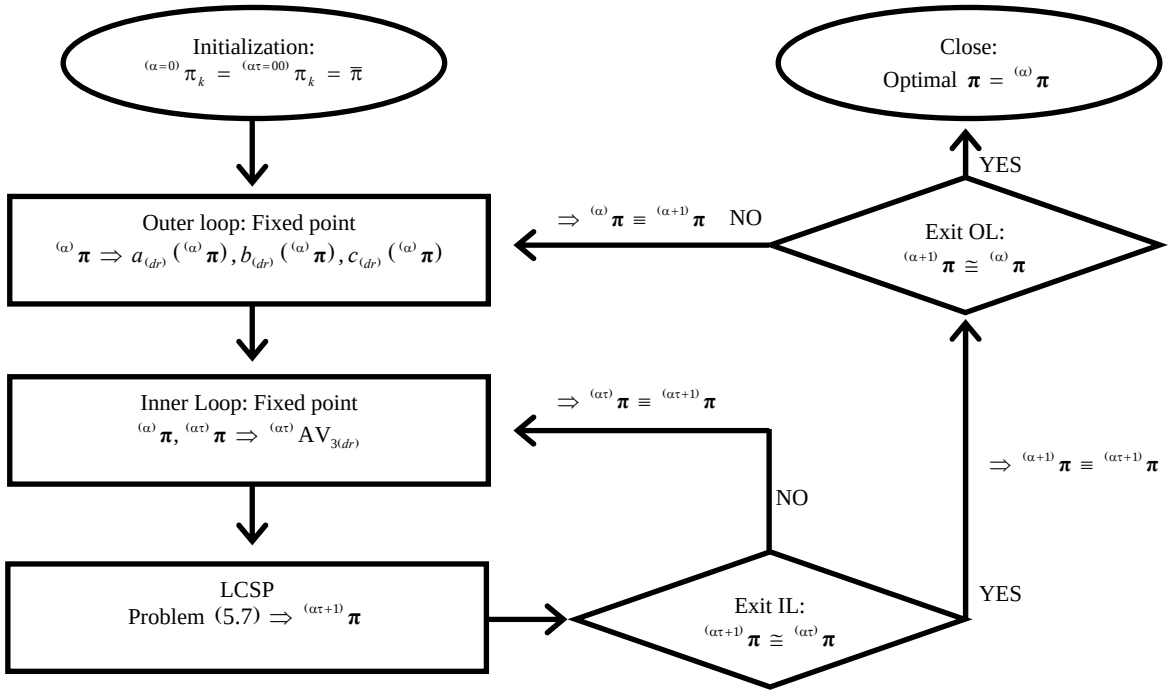


Figure 5.1 Algorithm flowchart

Initialization. At iteration $\alpha = 0$ of the OL, set $^{(\alpha=0)}\boldsymbol{\pi} = \left\{^{(\alpha=0)}\pi_k = \overline{\pi}; k = 1, \ldots, N\right\}$ with $0 < \overline{\pi} \leq 1$. A reasonable choice is $\overline{\pi} = 0.5$. At iteration $\tau = 0$ of the Inner Loop, set $^{(\alpha\tau=0)}\boldsymbol{\pi} = {}^{(\alpha)}\boldsymbol{\pi}$. Fix the N vector, $\boldsymbol{\varepsilon}$, of small positive values.

Outer loop

- ***Fixing the values for the Inner Loop***. In accordance with expressions (A1.4), (A1.7) and (A1.8) given in Appendix A1, the following real scalar values are computed

$$a_{(dr)k}(^{(\alpha)}\boldsymbol{\pi}) = \boldsymbol{\delta}_k' [\mathbf{A}(^{(\alpha)}\boldsymbol{\pi})]^{-1} \sum_{j \in U} \boldsymbol{\delta}_j \tilde{y}_{rj} \gamma_{dj} \left(1 - {}^{(\alpha)}\pi_j\right), \tag{5.3}$$

$$b_{(dr)k}(^{(\alpha)}\boldsymbol{\pi}) = \boldsymbol{\delta}_k' [\mathbf{A}(^{(\alpha)}\boldsymbol{\pi})]^{-1} \boldsymbol{\delta}_k \sigma_{rk}^2 \gamma_{dk} \left(1 - {}^{(\alpha)}\pi_k\right), \tag{5.4}$$

$$c_{(dr)k}(^{(\alpha)}\boldsymbol{\pi}) = \pi_k^2 \boldsymbol{\delta}_k' [\mathbf{A}(^{(\alpha)}\boldsymbol{\pi})]^{-1} \left[\sum_{j \in U} \boldsymbol{\delta}_j \boldsymbol{\delta}_j' \sigma_{rj}^2 \gamma_{dj} \left(1 - {}^{(\alpha)}\pi_j\right)^2\right] [\mathbf{A}(^{(\alpha)}\boldsymbol{\pi})]^{-1} \boldsymbol{\delta}_k. \tag{5.5}$$

- ***Launch of the Inner Loop***. The Inner Loop is executed until convergence.
- ***Updating or exiting.*** If the vector $^{(\alpha+1)}\boldsymbol{\pi}$ is such that $\left|^{(\alpha+1)}\boldsymbol{\pi} - {}^{(\alpha)}\boldsymbol{\pi}\right| > \boldsymbol{\varepsilon}$, then the Outer Loop is iterated by updating the vector $^{(\alpha)}\boldsymbol{\pi}$ with $^{(\alpha+1)}\boldsymbol{\pi}$. If $\left|^{(\alpha+1)}\boldsymbol{\pi} - {}^{(\alpha)}\boldsymbol{\pi}\right| \leq \boldsymbol{\varepsilon}$, then the Outer Loop closes and $^{(\alpha)}\boldsymbol{\pi}$ represents the optimal values solution to the problem of the system (5.1).

Inner Loop

- ***Fixing the values for the LCSP***. The following values are computed:

$$\begin{aligned} {}^{(\alpha\tau)}\mathrm{AAV}_{3(dr)} &= \sum_{k\in U}\left(1-{}^{(\alpha\tau)}\pi_k\right)a_{(dr)k}\left({}^{(\alpha)}\boldsymbol{\pi}\right)\left[2\tilde{y}_{rk}\gamma_{dk}-{}^{(\alpha\tau)}\pi_k a_{(dr)k}\left({}^{(\alpha)}\boldsymbol{\pi}\right)\right] \\ &+ \sum_{k\in U}\left(1-{}^{(\alpha\tau)}\pi_k\right)\left[2b_{(dr)k}\left({}^{(\alpha)}\boldsymbol{\pi}\right)-{}^{(\alpha\tau)}\pi_k c_{(dr)k}\left({}^{(\alpha)}\boldsymbol{\pi}\right)\right]. \end{aligned} \tag{5.6}$$

  in accordance with expression (A1.7) in Appendix A1.

- ***Solving the LCSP***. Considering the ${}^{(a\tau)}\mathrm{AAV}_{3(dr)}$ values as fixed, the ${}^{(\alpha\tau+1)}\boldsymbol{\pi}$ is obtained by solving, by a standard algorithm for a classical LCSP, the following optimization problem:

$$\begin{cases} \mathrm{Min}\left(\sum_{k\in U}{}^{(\alpha\tau+1)}\pi_k c_k\right) \\ \sum_{k\in U}\dfrac{\left(\tilde{y}_{rk}^2+\sigma_{rk}^2\right)\gamma_{dk}}{{}^{(\alpha\tau+1)}\pi_k}\le\dfrac{N-H}{N}\bar{V}_{(dr)}+\sum_{k\in U}\left(\tilde{y}_{rk}^2+\sigma_{rk}^2\right)\gamma_{dk}+{}^{(\alpha\tau)}\mathrm{AAV}_{3(dr)}. \\ 0<{}^{(\alpha\tau+1)}\pi_k\le 1 \qquad (k=1,\ldots,N) \end{cases} \tag{5.7}$$

- ***Updating or exiting.*** If the vector ${}^{(\alpha\tau+1)}\boldsymbol{\pi}$ is such that $\left|{}^{(\alpha\tau+1)}\boldsymbol{\pi}-{}^{(\alpha\tau)}\boldsymbol{\pi}\right|>\boldsymbol{\varepsilon}$, then the Inner Loop is iterated by updating the vector ${}^{(\alpha\tau)}\boldsymbol{\pi}$ with ${}^{(\alpha\tau+1)}\boldsymbol{\pi}$. If $\left|{}^{(\alpha\tau+1)}\boldsymbol{\pi}-{}^{(\alpha\tau)}\boldsymbol{\pi}\right|\le\boldsymbol{\varepsilon}$ then the Inner Loop closes and the updated vector ${}^{(\alpha+1)}\boldsymbol{\pi}$ for the Outer Loop is given by ${}^{(\alpha\tau+1)}\boldsymbol{\pi}$.

Remark 5.1. The problem of the system (5.7) can be solved by the algorithm proposed in Falorsi and Righi (2008, Section 3.1) which represents a slight modification of Chromy's algorithm (1987), originally developed for multivariate optimal allocation in SSRSWOR designs and implemented in standard software tools (see for example the Mauss-R software available at: http://www3.istat.it/strumenti/metodi/software/campione/mauss_r/). Alternatively, the LCSP can be dealt with by the SAS procedure NLP as suggested by Choudhry et al. (2012).

Remark 5.2. The algorithm distinguishes the ${}^{(\alpha)}\pi_k$ (updated in the Outer loop) from the ${}^{(\alpha\tau)}\pi_k$ (updated in the Inner loop). The innovation of the proposed algorithm lies precisely in this peculiarity. If this distinction between the inclusion probabilities is not made, i.e., ${}^{(\alpha\tau)}\boldsymbol{\pi}={}^{(\alpha)}\boldsymbol{\pi}$, we have observed in several experiments that the iterate solutions of the LCSP for each Outer Loop do not converge to a stationary point.

Remark 5.3. After the optimization phase, in which the $\boldsymbol{\pi}$ vector is defined as solution to problem of system (5.1), a *calibration phase* is performed (Falorsi and Righi 2008) to obtain calibrated inclusion probabilities, ${}_{\mathrm{cal}}\pi_k$, which modifies the *optimal* $\boldsymbol{\pi}$ vector marginally in order to satisfy $\sum_{k\in U}{}_{\mathrm{cal}}\pi_k\boldsymbol{\delta}_k=\mathbf{n}$, where $\mathbf{n}$ is a vector of integer numbers. The use of the Generalized Iterative Proportional Fitting algorithm (Dykstra and Wollan 1987) ensures that all resulting calibrated inclusion probabilities are in the $(0,1]$ interval.

6 Empirical evaluations

Several simulations were carried out on real and simulated data sets to investigate the empirical properties of the proposed sampling strategy. Here, we show the results obtained for a single real data exercise, referred to the 1999 population of enterprises having a number of employed persons between 1 and 99, and belonging to Computer and related economic activities (2-digits of the *Statistical classification of economic activities in the European Community rev.1*, abbreviated as NACE). Three experiments were performed. Experiment (a) checked whether the allocation obtained by the proposed algorithm converged towards the solution of the standard Chromy's algorithm for the SSRSWOR design. Experiment (b) compared the sample sizes of the standard SSRSWOR design with the Incomplete Stratified Sampling (ISS) design, in which the cross-classified strata were unplanned subpopulations; this experiment studied the risk of statistical burden due to repeated selection on different survey occasions. Finally, Experiment (c) measured the discrepancies between the expected Coefficients of Variation (CV) computed by the algorithm and the empirical CV obtained by a Monte Carlo simulation.

The c_k values were, in all three experiments, uniformly set equal to 1. The Anticipated Variance according to the approximation proposed in Remark 4.1 was also calculated.

The population chosen for the experiments had a size of $N = 10{,}392$ enterprises. The domains of interest identify two partitions of the target population: the *geographical region*, with 20 marginal domains (DOM1), and the *economic activity group* (3-digits of the NACE with 6 different groups) *by size class* (defined in terms of number of employed persons: $1 = 1-4; 2 = 5-9; 3 = 10-19;$ $4 = 20-99$), with 24 marginal domains (DOM2). The overall number of marginal domains was 44, while the number of cross-classified or multi-way strata with a not-zero population size was 360. The modal value of the population size distribution is 1, and 29.17% of the cross-classified strata have at most 2 units. This type of strata represents a critical issue in the context of standard stratified approaches. Indeed, for calculating unbiased variance estimates, these strata must be take-all strata (so that they do not contribute to the variance of the estimates), although the allocation rule would require fewer units and, in general, a non-integer number of sample units. The variables of interest were the *labour cost* and the *value added*, which are available for each population unit from an administrative data source. Typically both variables have highly skewed distributions.

The target estimates for all the empirical studies are the 88 totals at the domain level (2 variables by 44 marginal domains). In each experiment, the inclusion probabilities were determined by fixing the $\bar{V}_{(dr)} = (0.1 t_{(dr)})^2$ in (5.1), which is equivalent to fixing the maximum accepted level of the percent CV of the domain level estimates at 10%.

Empirical study (a). The first experiment took into account the partition DOM1. These domains represented both *planned* domains and *estimation* domains. Since the planned domains defined a partition of the population of interest, they could also be considered as strata in the standard sampling designs. The predictive working model was given by

$$\begin{cases} y_{rk} = \alpha_d + u_{rk} \;\; \forall k \in U_d \; (d = 1, \ldots, 20) \\ \\ E_M(u_{rk}) = 0, \; E_M(u_{rk}^2) = \sigma_{rd}^2 \;\; \forall k \in U_d; \;\; E_M(u_{rk}, u_{rl}) = 0 \;\; \forall k \neq l \end{cases}, \qquad (6.1)$$

where α_d is a fixed effect and the superpopulation variances σ^2_{rd} were estimated by means of the residual variance of the predictive model in each region. The algorithm proposed in Section 5 was performed using three different initial values of the inclusion probabilities $\bar{\pi}$, equal to 0.01, 0.50 and 0.99 respectively. The initial inclusion probability values had no impact on the final solution, although it was achieved with a different number of iterations. We note that the overall number of inner loops was 17 for $\bar{\pi} = 0.01$. The convergence was achieved with 13 inner loops for $\bar{\pi} = 0.50$; 14 inner loops were needed for $\bar{\pi} = 0.99$. However, after the ninth iteration, the three sampling sizes were quite similar (Figure 6.1). In the experiment, the overall sample sizes were 3,105 for the benchmark Chromy allocation and 3,110 for the method proposed here. However, the differences between the two sampling sizes at the domain level were fractional numbers that were always lower than 1, and with the absolute largest relative difference lower than 0.3%. This highlights that the proposed algorithm actually defines the same domain sampling sizes of those calculated by the benchmark allocation. With regards to convergence, the initial inclusion probability values have no impact on the final solution, although this is achieved with a different number of iterations.

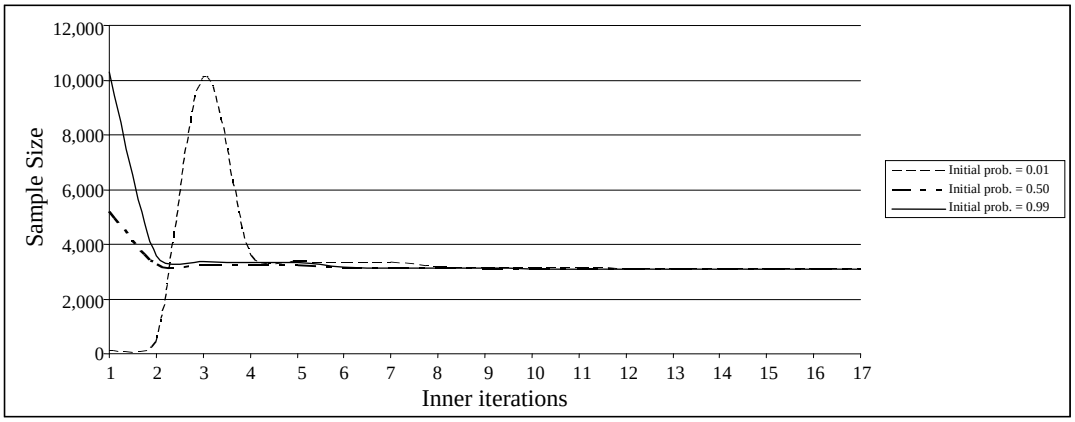


Figure 6.1 Convergence of the algorithm with different initial inclusion probabilities in the empirical study (a)

Similar results were obtained if the domains of interests were identified by the partition DOM2.

Empirical study (b). Let U_{d_1} be a specific region $(d_1 = 1, \ldots, 20)$ of DOM1, and let U_{d_2} (with $d_2 = 1, \ldots, 24$) be a specific economic activity group by the enterprise size class of the partition DOM2. Two prediction models, M_1 and M_2, were used. Referring to the notation of the ANOVA models, M_1 is the saturated model given by

$$\begin{cases} y_{rk} = \alpha_{d_1} + \lambda_{d_2} + (\alpha\lambda)_{d_1 d_2} + u_{rk} \;\; \forall k \in U_{d_1} \cap U_{d_2} \\ E_M(u_{rk}) = 0,\; E_M(u^2_{rk}) = \sigma^2_{r(d_1 d_2)} \;\; \forall k \in U_{d_1} \cap U_{d_2};\;\; E_M(u_{rk}, u_{rl}) = 0 \;\; \forall k \neq l \end{cases}, \qquad (6.2)$$

in which α_{d_1} and λ_{d_2} are the main effects, related to the domains U_{d_1} and U_{d_2} respectively and with $(\alpha\lambda)_{d_1 d_2}$ as the interaction effect. The model variances $\sigma^2_{r(d_1 d_2)}$ were estimated by means of the ordinary

least square method, by computing the variances of the residual terms at the $U_{d_1} \cap U_{d_2}$ level. Model M_2 is identical to model M_1 without the interaction factor. Table 6.1 shows the goodness of fit of the two models.

Table 6.1
Goodness of fit of the models used for the prediction

Model	Goodness of fit R^2%	
	Labour cost	Value added
Model M_1 (Expression 6.2)	68.1	64.1
Model M_2 (Expression 6.2 without interactions)	65.1	61.0

Three different allocations were considered for the SSRSWOR in the case of model M_1 : (*i*) no stratum sample size constraint is given; (*ii*) at least 1 sample unit per stratum is required (to obtain unbiased point estimates); (*iii*) at least 2 sample units per stratum are required (to achieve unbiased variance estimates) for all strata having a population size of 2 or more enterprises. The first two allocations were rather theoretical since in all the business surveys conducted by the Italian National Statistical Institute, the selection of at least two units per stratum is required. The results of the experiment are shown in Table 6.2 below. Only the results for the case in which the initial inclusion probabilities were equal to $\bar{\pi} = 0.50$ are investigated herein; identical sample sizes were obtained with the other initial values of the inclusion probabilities, with a slightly slower convergence process. The three SSRSWOR designs have 716.6, 944 and 1,042 sample units respectively. The Incomplete stratified Sampling (ISS) design with model M_1 led to 936 units; while model M_2 led to 991 units. The better result obtained by model M_1 with respect to model M_2 was due to the fact that model M_1 had a better fit. Finally, the ISS designs helped tackling the statistical burden of respondent enterprises. Indeed, assuming that the inclusion probabilities remain fixed for the different survey occasions, their distributions may be used to assess the statistical burden in repeated surveys. Table 6.2 shows that the number of enterprises drawn with certainty in each survey occasion was 175 for the third SSRSWOR designs, while 30 and 40 enterprises were selected with certainty in the first and second ISS designs, respectively. Analysing the sizes (in terms of employed persons) of the enterprises included in the sample with certainty, the third SSRSWOR design had an average size equal to 20.6. In some cases, enterprises with 2 employed persons were included in the sample with certainty. Conversely, we observe that in the first and second ISS designs, the enterprises with minimum size had 17 and 16 employed persons respectively, and an average size larger than 40 units.

Table 6.2
Sample sizes and distribution of the enterprises included in the sample with certainty, for different sampling designs

Sampling design		Sample size	Enterprises selected with certainty		
			Number	Number of employed	
				Average	Minimum
Standard Stratified with M_1 model	No stratum sample size constraint	716.6	10	47.0	23.0
	At least 1 sample unit per stratum	944.0	119	24.0	2.0
	At least 2 sample units per stratum	1,042.0	175	20.6	2.0
Incomplete Stratified Sampling with M_1 model		936.0	30	50.1	17.0
Incomplete Stratified Sampling with M_2 model without interactions		991.0	40	42.9	16.0

Finally, to assess the solution's sensitivity, the experiment was repeated artificially and the prediction values of $\tilde{y}_{rk}$ and $\tilde{\sigma}^2_{rk}$ in the optimization problem (5.1) were changed. In particular, we increased the prediction values of $\tilde{\sigma}^2_{rk}$ by 20% and 120% respectively, and decreased by 20% the $\tilde{y}_{rk}$ values predicted by model M_1. As expected, the sample sizes increased, but the SSRSWOR design with at least 1 sample unit per stratum and the first ISS design roughly defined the same sample sizes (Table 6.3).

Table 6.3
Sample sizes with modified expected values of the predictions of model (4.1)

Sampling design		Sample size		
		$\tilde{\sigma}^2_{rk}$ increased by 20%	$\tilde{\sigma}^2_{rk}$ increased by 120%	$\tilde{y}_{rk}$ decreased by 20%
SSRSWOR with M_1 model	No stratum sample size constraint	821.0	1,269.0	993.8
	At least 1 sample unit per stratum	1,035.0	1,472.0	1,206.0
	At least 2 sample units per stratum	1,125.0	1,536.0	1,283.0
ISS design with M_1 model		1,039.7	1,460.9	1,207.5

Empirical study (c). The heteroschedastic linear prediction model M_3 was used:

$$\begin{cases} y_{rk} = \alpha_r + \varphi_r x_k + u_{rk} \\ E_M(u_{rk}) = 0, \;\; E_M(u^2_{rk}) = \sigma^2_{rk} = \sigma^2_r x_k \;\; \forall k \in U; \; E_M(\varepsilon_{rk}, \varepsilon_{rl}) = 0 \;\; \forall k \neq l \end{cases}, \tag{6.3}$$

where x_k is the *number of employed persons* in the k^{th} enterprise, and α_r and φ_r are the regression parameters. Note that the number of employed persons is available in the sampling frame in Italy.

Two different model variance estimates were carried out:
(a) $\tilde{\sigma}^2_{rk} = 1/N_{(X=x_k)} \sum_{k \in U_{(X=x_k)}} (y_{rk} - \mathrm{A}_r - \mathrm{F}_r x_k)^2$ and (b) $\tilde{\sigma}^2_{rk} = \tilde{\sigma}^2_r x_k$, in which $\tilde{\sigma}^2_r = 1/(N-2) \sum_{k \in U} [(y_{rk} - \mathrm{A}_r - \mathrm{F}_r x_k)/x_k]^2$, where $U_{(X=x)}$ is the population of enterprises, of size $N_{(X=x)}$, for which the variable X assumes the value x; A_r and F_r are the weighted least square estimates for the complete enumerated population of α_r and φ_r respectively. The sum of the estimated model variances obtained with method (a) is smaller than that obtained with method (b). This was reflected in the computed sample sizes. The first allocation defined an overall sample size of 927 units, while the sample size of the second allocation was 951. Successively, 1,000 samples were drawn for both allocations and the ratios $\mathrm{RCV}(\hat{t}_{(dr)}) = \mathrm{ECV}(\hat{t}_{(dr)})/\mathrm{SCV}(\hat{t}_{(dr)})$ were calculated, with $\mathrm{ECV}(\hat{t}_{(dr)}) = \left[\sqrt{\mathrm{AAV}(\hat{t}_{(dr)})}/\hat{t}_{(dr)}\right]100$ as the expected CV (%) and

$$\mathrm{SCV}(\hat{t}_{(dr)}) = 100 \sqrt{(1/I)\left[\sum_{i=1}^{I} \hat{t}^i_{(dr)} - (1/I)\sum_{i=1}^{I} \hat{t}^i_{(dr)}\right]^2} \Big/ (1/I)\sum_{i=1}^{I} \hat{t}^i_{(dr)}$$

as the simulated (or empirical) CV, obtained as a result of the simulation, having denoted with $\hat{t}^i_{(dr)}$ the HT estimate in the i^{th} iteration and $I = 1{,}000$. For the sake of brevity, only the the main results of allocation (b) are shown in Figure 6.2, for DOM1 and DOM2 respectively, and both variables of interest. Examining the figure on the left, we emphasize that the simulation generally produces a simulated CV that is smaller

than expected, with an RCV ratio larger than 1 for both variables. One exception occurs, for the value added in one domain of DOM1.

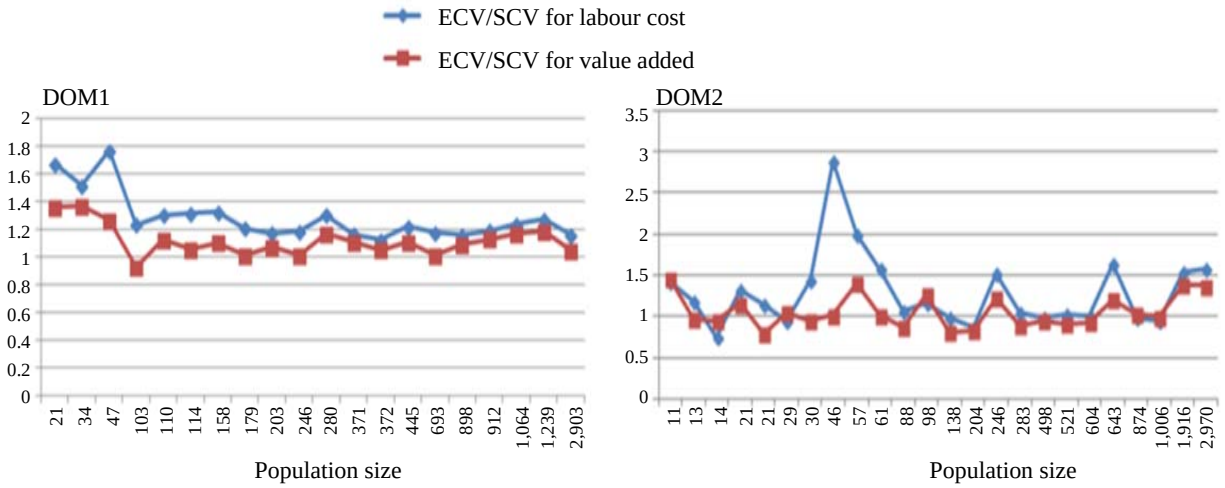


Figure 6.2 RCVs by population size for labour cost and value added

RCV lower than 1 may be explained by the increase of the domain sample sizes, due to the calibration step. We note that in general, these discrepancies are observed in domains with a small population size; thus, the calibration step may have a non-negligible impact. The figure on the right shows more articulated and conflicting empirical evidence. First, we note that the RCV are often larger or very close to 1. Nevertheless, in three domains, the value added variable has simulated CV's equal to 11.5%, 12.0% and 12.3%. In these rare cases, and in some others (labour cost in two domains), the discrepancies are coherent with the findings of Deville and Tillé (2005) on the empirical properties of variance approximation for balanced sampling.

7 Conclusions

The paper proposes a new approach for defining the optimal inclusion probabilities in various survey contexts, which are characterized by the need to disseminate survey estimates of prefixed accuracy, for a multiplicity of both variables and domains of interest.

This paper's main contribution is the practical computation of these probabilities by means of a new algorithm, which is suitable for a general multi-way sampling design in which the standard stratified sampling represents a special case. The proposed approach, the algorithm and the final computation are domain- and variable-driven.

In our framework, the domain membership indicator variables are assumed to be known, while the variables of interest are not known. The procedure is, then, applied on the predicted values of the characteristics of interest via a superpopulation model, and the algorithm enables taking into account

model uncertainty; this reflects the non-knowledge of the values of variables of interest. Using the Anticipated Variance as the measure of the estimators' precision, this approach overcomes the limits of the standard algorithms for the sample allocation, in which the variables of interest driving the solution are assumed to be known.

The proposed algorithm exploits standard procedure, but does present some computational innovations which may be useful for dealing with the complexity deriving from the fact that the Anticipated Variances are implicit functions of the inclusion probabilities. The algorithm was tested on simulated and real survey data, to evaluate its performance and properties. The results of a small set of experiments are presented here. They confirm an improvement, in terms of efficiency, of the sampling strategy. A natural generalization of the case examined here may be developed by considering, as known during the design planning stage, the indicators of the domains and other quantitative independent variables. We note that the Anticipated Variance considering only the domain indicators is larger than the Anticipated Variance of this more general case. Thus, our solution represents an upper (and somehow robust) boundary solution in the design phase. Furthermore, the algorithmic solution can be easily adapted to this more general situation.

Acknowledgements

This research was funded by the partnership of the Global Strategy to improve Agricultural and Rural Statistics: http://www.fao.org/economic/ess/ess-capacity/ess-strategy/en/.

Appendix A1

AV of the HT estimator

Let us consider the residual $\eta_{(dr)k}$ as expressed by equation (3.5), and replace the term y_{rk} with $\tilde{y}_{rk} + u_{rk}$, thus obtaining

$$\eta_{(dr)k} = (\tilde{y}_{rk} + u_{rk})\gamma_{dk} - \pi_k \boldsymbol{\delta}'_k [\mathbf{A}(\boldsymbol{\pi})]^{-1} \sum_{j \in U} \pi_j \boldsymbol{\delta}_j (\tilde{y}_{rj} + u_{rj})\gamma_{dj} (1/\pi_j - 1). \quad \text{(A1.1)}$$

The weighted least predictions of $\tilde{y}_{rk}\gamma_{dk}$ and $u_{rk}\gamma_{dk}$, with predictors $\pi_k \boldsymbol{\delta}_k$ and weights $1/\pi_k - 1$, are

$$\hat{\tilde{y}}_{(dr)k} = \pi_k a_{(dr)k} \quad \text{(A1.2)}$$

and

$$\hat{u}_{(dr)k} = \pi_k \boldsymbol{\delta}'_k [\mathbf{A}(\boldsymbol{\pi})]^{-1} \sum_{j \in U} \pi_j \boldsymbol{\delta}_j u_{rj}\gamma_{dj} (1/\pi_j - 1), \quad \text{(A1.3)}$$

with

$$a_{(dr)k}(\boldsymbol{\pi}) = \boldsymbol{\delta}'_k [\mathbf{A}(\boldsymbol{\pi})]^{-1} \sum_{j \in U} \pi_j \boldsymbol{\delta}_j \tilde{y}_{rj}\gamma_{dk} (1/\pi_j - 1). \quad \text{(A1.4)}$$

Using the formulae (A1.2) and (A1.3), the expression (A1.1) may be reformulated as $\eta_{(dr)k} = (\tilde{y}_{rk} + u_{rk})\gamma_{dk} - [\hat{\tilde{y}}_{(dr)k} + \hat{u}_{(dr)k}]$. Therefore, the model expectation of $\eta^2_{(dr)k}$ is

$$E_M(\eta^2_{(dr)k}) = (\tilde{y}_{rk}\gamma_{dk} - \hat{\tilde{y}}_{(dr)k})^2 + E_M[(u_{rk}\gamma_{dk} - \hat{u}_{(dr)k})^2] + \text{Mean zero terms}, \quad \text{(A1.5)}$$

because $E_M(u_{rk}) = 0$. Furthermore,

$$E_M[(u_{rk}\gamma_{dk} - \hat{u}_{(dr)k})^2] = \sigma^2_{rk}\gamma_{dk} + E_M(\hat{u}_{(dr)k})^2 - 2E_M(u_{rk}\gamma_{dk}, \hat{u}_{(dr)k}), \quad \text{(A1.6)}$$

where $E_M(u_{rk}\gamma_{dk}\hat{u}_{(dr)k}) = \pi_k b_{(dr)k}(\boldsymbol{\pi})$ and $E_M(\hat{u}_{(dr)k})^2 = \pi^2_k c_{(dr)k}(\boldsymbol{\pi})$, with

$$b_{(dr)k}(\boldsymbol{\pi}) = \boldsymbol{\delta}'_k[\mathbf{A}(\boldsymbol{\pi})]^{-1}\boldsymbol{\delta}_k\sigma^2_{rk}\gamma_{dk}(1-\pi_k) \quad \text{(A1.7)}$$

and

$$c_{(dr)k}(\boldsymbol{\pi}) = \boldsymbol{\delta}'_k[\mathbf{A}(\boldsymbol{\pi})]^{-1}\left[\sum_{j\in U}\boldsymbol{\delta}_j\boldsymbol{\delta}'_j\sigma^2_{rj}\gamma_{dj}(1-\pi_j)^2\right][\mathbf{A}(\boldsymbol{\pi})]^{-1}\boldsymbol{\delta}_k. \quad \text{(A1.8)}$$

Expression (4.5) is easily derived by plugging expressions from (A1.2) to (A1.8) into equation (4.3).

Appendix A2

Convergence of the algorithm

The optimization problem (5.1) is solved by two nested *fixed point iterations*. Given an unknown vector $\mathbf{x}$ of dimension q, the fixed point iteration chooses an initial guess ${}^{(0)}\mathbf{x}$. Then, it computes subsequent iterates by ${}^{(\tau+1)}\mathbf{x} = \mathbf{g}({}^{(\tau)}\mathbf{x})$, with $\tau = 1, 2, \ldots$, with $\mathbf{g}(\cdot)$ being a system of q updating equations. The multivariate function $\mathbf{g}$ has a fixed point in a domain $Q \subseteq \Re^q$ if $\mathbf{g}$ maps Q in Q. Let $J_{\mathbf{g}}(\mathbf{x})$ be the Jacobian matrix of first partial derivate of $\mathbf{g}$ evaluated at $\mathbf{x}$, if there exists a constant $\rho < 1$ such that, in some natural matrix norm, $\|J_{\mathbf{g}}(\mathbf{x})\| \leq \rho, \mathbf{x} \in Q$, $\mathbf{g}$ has a unique fixed point $\mathbf{x}^* \in Q$, and the fixed point iteration is guaranteed to converge to $\mathbf{x}^*$ for any initial guess chosen in Q. As regards the proposed algorithm, the convergence of the IL and OL is obtained when the terms ${}^{(\alpha\tau)}\mathrm{AAV}_{3(dr)}$ converge to the fixed point. This means that the vectors ${}^{(\alpha)}\boldsymbol{\pi}$ and ${}^{(\alpha\tau)}\boldsymbol{\pi}$ do not change in the OL and IL iterations. The demonstration below considers the method proposed by Chromy (1987) to solve the LCSP of system (5.7), and makes use of some reasonable assumptions: (1) $\hat{u}_{(dr)k} \cong 0$; (2) $[N/(N-H)] \cong 1$; (3) $\hat{\tilde{y}}_{rk} \cong \tilde{y}_{rk}$; (4) ${}^{(\alpha)}\pi_k \cong {}^{(\alpha\tau)}\Delta\,{}^{(\alpha\tau)}\pi_k$ with $0 < {}^{(\alpha\tau)}\Delta \leq 1$; (5) $c_k \cong \bar{c}$. Assumption (1) corresponds to the upward approximation of the Anticipated Variance, given in Remark 4.1, and implies that $b_{(dr)k}({}^{(\alpha)}\boldsymbol{\pi}) = c_{(dr)k}({}^{(\alpha)}\boldsymbol{\pi}) = 0$. Assumption (3) implies that $a_{(dr)k}({}^{(\alpha)}\boldsymbol{\pi})\tilde{y}_{rk}\gamma_{dk} \cong \tilde{y}^2_{rk}\gamma_{dk}/{}^{(\alpha)}\pi_k$. Assumption (4) states that the structure of the inclusion probabilities remains roughly constant in the different IL iterations. The assumption becomes reasonable considering that the updating equation A2.2 below (of a given inclusion probability) is essentially determined by the variance threshold that requires the largest sample size. It is plausible to hypothesize that this threshold remains more or less the same in the subsequent IL iterations of a given OL.

Proof of convergence of the Inner Loop. By reformulating expression (4.6) in accordance with the assumptions from (1) to (4),

$$ {}^{(\alpha\tau+1)}\mathrm{AAV}_{3(dr)} = \sum_{k\in U}\left[\left(\frac{1}{{}^{(\alpha\tau+1)}\pi_k}-1\right)\left(2\frac{\tilde{y}_{rk}^2\gamma_{dk}}{{}^{(\alpha\tau+1)}\Delta}-\frac{\tilde{y}_{rk}^2\gamma_{dk}}{{}^{(\alpha\tau+1)}\Delta^2}\right)\right]. \tag{A2.1} $$

Considering in problem (5.7) that the ${}^{(a\tau)}\mathrm{AAV}_{3(dr)}$ values are fixed, each value of the vector ${}^{(\alpha\tau+1)}\boldsymbol{\pi}$ is obtained as a solution of the LCSP with the Chromy algorithm. Denote with $\alpha\tau v^*$ the iteration of the Chromy algorithm into which it converges, where ${}^{(\alpha\tau v^*+1)}\boldsymbol{\pi} \cong {}^{(\alpha\tau v^*)}\boldsymbol{\pi}$. Then, the IL updates the generic probability in accordance with the expression

$$ {}^{(\alpha\tau+1)}\pi_k = \left[\sum_{(dr)} {}^{(\alpha\tau v^*+1)}\phi_{(dr)}\frac{\left(\tilde{y}_{rk}^2+\sigma_{rk}^2\right)\gamma_{dk}}{\overline{c}}\right]^{1/2}, \tag{A2.2} $$

where the right-hand term represents the updating formula of the Chromy algorithm, and $\sum_{(dr)}$ stands for $\sum_{d=1}^{D}\sum_{r=1}^{R}$, and ${}^{(\alpha\tau v^*+1)}\phi_{(dr)}$ is the generalized Lagrange multiplier, where

$$ \begin{aligned} {}^{(\alpha\tau v^*+1)}\phi_{(dr)} &= {}^{(\alpha\tau v^*)}\phi_{(dr)}\left[\frac{{}^{(\alpha\tau v^*)}V_{(dr)}}{\ddot{V}_{(dr)}+{}^{(\alpha\tau)}\mathrm{AAV}_{3(dr)}}\right]^2, \\ {}^{(\alpha\tau v^*)}V_{(dr)} &= \sum_{k\in U}\frac{\left(\tilde{y}_{rk}^2+\sigma_{rk}^2\right)\gamma_{dk}}{{}^{(\alpha\tau v^*)}\pi_k} \end{aligned} \tag{A2.3} $$

and

$$ \ddot{V}_{(dr)} = \overline{V}_{(dr)} + \sum_{k\in U}\left(\tilde{y}_{rk}^2+\sigma_{rk}^2\right)\gamma_{dk}. $$

The Kuhn-Tucker theory states that ${}^{(\alpha\tau v^*)}\phi_{(dr)}\left[{}^{(\alpha\tau v^*)}V_{(dr)} - \left(\ddot{V}_{(dr)} + {}^{(\alpha\tau)}\mathrm{AV}_{3(dr)}\right)\right] = 0$; therefore, ${}^{(\alpha\tau v^*+1)}\phi_{(dr)} = {}^{(\alpha\tau v^*)}\phi_{(dr)}$ and ${}^{(\alpha\tau v^*+1)}\phi_{(dr)} > 0$ iff ${}^{(\alpha\tau v^*)}V_{(dr)}\big/\left(\ddot{V}_{(dr)} + {}^{(\alpha\tau)}\mathrm{AV}_{3(dr)}\right) = 1$. Chromy asserts that few ${}^{(\alpha\tau v^*)}\phi_{(dr)}$ (for $r = 1,\ldots,R; d = 1,\ldots,D$) are larger than zero, and that in most cases, only one value is strictly positive. Denoting with ${}^{(\alpha\tau)}\mathbf{AAV}_3 = \left({}^{(\alpha\tau)}\mathrm{AAV}_{3(11)},\ldots,{}^{(\alpha\tau)}\mathrm{AAV}_{3(1R)},\ldots,{}^{(\alpha\tau)}\mathrm{AAV}_{3(DR)}\right)'$, we define ${}^{(\alpha\tau+1)}\mathbf{AAV}_3 = \mathbf{g}\left({}^{(\alpha\tau)}\mathbf{AAV}_3\right)$ as the system of $D\times R$ updating equations where the generic $(\overline{dr})$ equation of the system

$$ \begin{aligned} g_{(\overline{dr})}\left({}^{(\alpha\tau)}\mathbf{AAV_3}\right) &\cong \sum_{k\in U}\left(2\frac{\tilde{y}_{\overline{r}k}^2\gamma_{\overline{d}k}}{{}^{(\alpha\tau+1)}\Delta}-\frac{\tilde{y}_{\overline{r}k}^2\gamma_{\overline{d}k}}{{}^{(\alpha\tau+1)}\Delta^2}\right) \\ &\times \left\{\left[\sum_{(dr)} {}^{(\alpha\tau v^*)}\phi_{(dr)}\left[\frac{{}^{(\alpha\tau v^*)}V_{(dr)}}{\ddot{V}_{(dr)}+{}^{(\alpha\tau)}\mathrm{AAV}_{3(dr)}}\right]^2\frac{\left(\tilde{y}_{rk}^2+\sigma_{rk}^2\right)\gamma_{dk}}{\overline{c}}\right]^{-1/2}-1\right\}, \end{aligned} \tag{A2.4} $$

is obtained by plugging expression (A2.2) into (A2.1). If the convergence is obtained, then in the last iteration, ${}^{(\alpha\tau+1)}\mathbf{AAV}_3 \cong {}^{(\alpha\tau)}\mathbf{AAV}_3$. The function of equation (A2.4) is continuous and differentiable. Moreover, it maps onto the interval of the possible values of $\mathrm{AAV}_{3(dr)}$. Then, the IL converges if the following condition is fulfilled:

$$\left\| J_{\mathbf{g}}\left(\mathbf{AAV}_3\right)\right\| \leq 1. \tag{A2.5}$$

The Jacobian matrix is positive semi-defined, and a well-known result states that trace $\left(J_{\mathbf{g}}J'_{\mathbf{g}}\right) \leq$ trace $\left(J_{\mathbf{g}}\right)^2$. By considering the Frobenius norm $\left\|J_{\mathbf{g}}\right\|_F = \sqrt{\text{trace}\left(J_{\mathbf{g}}J'_{\mathbf{g}}\right)}$, it is $\left\|J_{\mathbf{g}}\right\|_F \leq$ trace $\left(J_{\mathbf{g}}\right)$. Thus we can take into account the trace of the Jacobian matrix to verify condition (A2.5). Let $g'_{(\overline{dr})} = \partial g_{(\overline{dr})}\left({}^{(\alpha\tau-1)}\mathbf{AAV}_{3(dr)} \big/ \partial\, {}^{(\alpha\tau-1)}\mathrm{AAV}_{3(\overline{dr})}\right)$ be the $(\overline{dr})$ element of the diagonal of $J_{\mathbf{g}}\left(\mathbf{AAV}_3\right)$. Using the Kuhn-Tucker condition ${}^{(\alpha\tau v^*)}V_{(dr)} \big/ \left(\ddot{V}_{(dr)} + {}^{(\alpha\tau)}\mathrm{AV}_{3(dr)}\right) = 1$,

$$g'_{(\overline{dr})} = \sum_{k\in U}\left(2\frac{\tilde{y}^2_{\overline{r}k}\gamma_{\overline{d}k}}{{}^{(\alpha\tau+1)}\Delta} - \frac{\tilde{y}^2_{\overline{r}k}\gamma_{\overline{d}k}}{{}^{(\alpha\tau+1)}\Delta^2}\right)\left[\sum_{(dr)} {}^{(\alpha\tau v^*)}\phi_{(dr)}\frac{\left(\tilde{y}^2_{rk} + \sigma^2_{rk}\right)\gamma_{dk}}{\overline{c}}\right]^{-3/2}$$

$$\times \; {}^{(\alpha\tau v^*)}\phi_{(\overline{dr})}\frac{1}{{}^{(\alpha\tau v^*)}V_{(\overline{dr})}}\frac{\left(\tilde{y}^2_{\overline{r}k} + \sigma^2_{\overline{r}k}\right)\gamma_{\overline{d}k}}{\overline{c}}.$$

Since many ${}^{(\alpha\tau v^*)}\phi_{(\overline{dr})} = 0$ (Chromy 1987), the respective $g'_{(\overline{dr})}$ is null. When ${}^{(\alpha\tau v^*)}\phi_{(\overline{dr})} > 0$, then

$$g'_{(\overline{dr})} \leq \sum_{k\in U}\left(2\frac{\tilde{y}^2_{\overline{r}k}\gamma_{\overline{d}k}}{{}^{(\alpha\tau+1)}\Delta} - \frac{\tilde{y}^2_{\overline{r}k}\gamma_{\overline{d}k}}{{}^{(\alpha\tau+1)}\Delta^2}\right)\left[{}^{(\alpha\tau v^*)}\phi_{(\overline{dr})}\frac{\left(\tilde{y}^2_{\overline{r}k} + \sigma^2_{\overline{r}k}\right)\gamma_{\overline{d}k}}{\overline{c}}\right]^{-3/2} \times {}^{(\alpha\tau v^*)}\phi_{(\overline{dr})}\frac{1}{{}^{(\alpha\tau v^*)}V_{(\overline{dr})}}\frac{\left(\tilde{y}^2_{\overline{r}k} + \sigma^2_{\overline{r}k}\right)\gamma_{\overline{d}k}}{\overline{c}}$$

$$= \sum_{k\in U}\left(2\frac{\tilde{y}^2_{\overline{r}k}\gamma_{\overline{d}k}}{{}^{(\alpha\tau+1)}\Delta} - \frac{\tilde{y}^2_{\overline{r}k}\gamma_{\overline{d}k}}{{}^{(\alpha\tau+1)}\Delta^2}\right)\frac{1}{\sqrt{{}^{(\alpha\tau v^*)}\phi_{(\overline{dr})}\frac{\left(\tilde{y}^2_{\overline{r}k} + \sigma^2_{\overline{r}k}\right)\gamma_{\overline{d}k}}{\overline{c}}\,{}^{(\alpha\tau v^*)}V_{(\overline{dr})}}}$$

$$\leq \sum_{k\in U}\frac{\frac{\tilde{y}_{\overline{r}k}\gamma_{\overline{d}k}}{{}^{(\alpha\tau+1)}\Delta}\left(2 - \frac{1}{{}^{(\alpha\tau+1)}\Delta}\right)}{\sqrt{\overline{c}\,{}^{(\alpha\tau v^*)}\phi_{(\overline{dr})}\gamma_{\overline{d}k}}\,{}^{(\alpha\tau v^*)}V_{(\overline{dr})}} << 1.$$

Therefore, the trace $\left(J_{\mathbf{g}}\right)$ should be less than 1.

Proof of convergence of the Outer Loop. Let ${}^{(\alpha\tau+1)}\boldsymbol{\pi}$ be the fixed point solution of the IL; then, the OL updates the vector ${}^{(\alpha)}\boldsymbol{\pi}$ with ${}^{(\alpha+1)}\boldsymbol{\pi} = {}^{(\alpha\tau+1)}\boldsymbol{\pi}$. Under conditions (1), (2) and (3),

$${}^{(\alpha+1)}\mathrm{AAV}_{3(dr)} = \sum_{k\in U}\left(\frac{1}{{}^{(\alpha\tau+1)}\pi_k} - 1\right)\tilde{y}^2_{rk}\gamma_{dk}. \tag{A2.6}$$

Plugging expression (A2.2) into formula (A2.6) when the IL converges, the system of $D \times R$ updating equations of ${}^{(\alpha+1)}\mathbf{AAV}_3$ is given by ${}^{(\alpha+1)}\mathbf{AAV}_3 = \mathbf{j}\left({}^{(\alpha\tau)}\mathbf{AAV}_3\right)$, where the generic equation of $\mathbf{j}$ is

$$\begin{aligned} {}^{(\alpha+1)}\mathrm{AAV}_{3(dr)} &= j_{(\overline{dr})}\left({}^{(\alpha\tau)}\mathbf{AAV}_3\right) \\ &= \sum\nolimits_{k\in U} \tilde{y}_{\overline{r}k}^2 \gamma_{\overline{d}k} \left(\left[\sum\nolimits_{(dr)} {}^{(\alpha\tau\nu^*)}\phi_{(dr)} \left[\frac{{}^{(\alpha\tau\nu^*)}V_{(dr)}}{\ddot{V}_{(\overline{dr})} + {}^{(\alpha\tau)}\mathrm{AAV}_{3(\overline{dr})}} \right]^2 \frac{\left(\tilde{y}_{rk}^2 + \sigma_{rk}^2\right)\gamma_{dk}}{\overline{c}} \right]^{-1/2} - 1 \right). \end{aligned} \tag{A2.7}$$

Denoting with ${}^{(\alpha)}\mathbf{AAV}_3 = {}^{(\alpha\tau=0)}\mathbf{AAV}_3$, the system $\mathbf{j}$ may be expressed in a recursive form

$${}^{(\alpha+1)}\mathbf{AAV}_3 \cong \mathbf{j}\left(\mathbf{g}\left({}^{(\alpha\tau-1)}\mathbf{AAV}_3\right)\right) = \mathbf{j}\left(\mathbf{g}\left(\mathbf{g}\left(.....\mathbf{g}\left({}^{(\alpha\tau=0)}\mathbf{AAV}_3\right)\right)\right)\right) = \mathbf{f}\left({}^{(\alpha)}\mathbf{AAV}_3\right),$$

with $\mathbf{f}(\cdot) = \mathbf{j}(\mathbf{g}(\mathbf{g}(.....\mathbf{g}(\cdot))))$ as the system of $D \times R$ updating equations of ${}^{(\alpha+1)}\mathbf{AAV}_3$, with respect to the previous values of the OL, ${}^{(\alpha)}\mathbf{AAV}_3$. To demonstrate the convergence of OL, it is necessary to demonstrate that the Jacobian norm $\left\|J_{\mathbf{f}}\left(\mathbf{AAV}_3\right)\right\|$ is lower than 1. Using standard results of matrix algebra,

$$\left\|J_{\mathbf{f}}\left(\mathbf{AAV}_3\right)\right\| \le \left\|J_{\mathbf{j}}\left({}^{(\alpha\tau)}\mathbf{AAV}_3\right)\right\| \times \left\|J_{\mathbf{g}}\left({}^{(\alpha\tau-1)}\mathbf{AAV}_3\right)\right\| \times \ldots \times \left\|J_{\mathbf{g}}\left({}^{(\alpha\tau=0)}\mathbf{AAV}_3\right)\right\|,$$

in which the generic norm $\left\|J_{\mathbf{g}}(\cdot)\right\|$ is lesser than 1 (see the IL proof of convergence). Let $j'_{(\overline{dr})}$ be the $(\overline{dr})$ element on the diagonal of $J_{\mathbf{j}}\left({}^{(\alpha\tau)}\mathbf{AAV}_3\right)$. It is

$$\begin{aligned} j'_{(\overline{dr})} &= \sum\nolimits_{k\in U} \tilde{y}_{\overline{r}k}^2 \gamma_{\overline{d}k} \left[\sum\nolimits_{(dr)} {}^{(\alpha\tau\nu^*)}\phi_{(dr)} \frac{\left(\tilde{y}_{rk}^2 + \sigma_{rk}^2\right)\gamma_{dk}}{\overline{c}} \right]^{-3/2} \\ &\times {}^{(\alpha\tau\nu^*)}\phi_{(\overline{dr})} \frac{1}{{}^{(\alpha\tau\nu^*)}V_{(\overline{dr})}} \frac{\left(\tilde{y}_{\overline{r}k}^2 + \sigma_{\overline{r}k}^2\right)\gamma_{\overline{d}k}}{\overline{c}}. \end{aligned} \tag{A2.8}$$

Therefore, we have

$$\begin{aligned} j'_{(\overline{dr})} &\le \sum\nolimits_{k\in U} \tilde{y}_{\overline{r}k}^2 \gamma_{\overline{d}k} \left[{}^{(\alpha\tau\nu^*)}\phi_{(\overline{dr})} \frac{\left(\tilde{y}_{\overline{r}k}^2 + \sigma_{\overline{r}k}^2\right)\gamma_{\overline{d}k}}{\overline{c}} \right]^{-3/2} {}^{(\alpha\tau\nu^*)}\phi_{(\overline{dr})} \frac{1}{{}^{(\alpha\tau\nu^*)}V_{(\overline{dr})}} \frac{\left(\tilde{y}_{\overline{r}k}^2 + \sigma_{\overline{r}k}^2\right)\gamma_{\overline{d}k}}{\overline{c}} \\ &= \frac{1}{{}^{(\alpha\tau\nu^*)}V_{(\overline{dr})}} \sum\nolimits_{k\in U} \tilde{y}_{\overline{r}k}^2 \gamma_{\overline{d}k} \left[{}^{(\alpha\tau\nu^*)}\phi_{(\overline{dr})} \frac{\left(\tilde{y}_{\overline{r}k}^2 + \sigma_{\overline{r}k}^2\right)\gamma_{\overline{d}k}}{\overline{c}} \right]^{-1/2}. \end{aligned}$$

The following inequality holds

$$j'_{(\overline{dr})} < \frac{\sum_{k\in U} \tilde{y}_{\overline{r}k}\gamma_{\overline{d}k}}{\sqrt{\overline{c}\,{}^{(\alpha\tau\nu^*)}\phi_{(\overline{dr})}}\,{}^{(\alpha\tau\nu^*)}V_{(\overline{dr})}} << 1.$$

Consequently, the norm $\left\|J_{\mathbf{j}}\left({}^{(\alpha\tau)}\mathbf{AAV}_3\right)\right\| < 1$, and therefore the OL converges.

Appendix A3

Proof that the approximation of Remark 4.1 is upward

Since $\hat{u}_{(dr)k}$ is the weighted least square prediction of $u_{rk}\gamma_{dk}$, by using a different value of the $\hat{u}_{(dr)k}$, such as $\hat{u}_{(dr)k} = 0$, we obtain

$$\sum_{k\in U}(1/\pi_k - 1)\, E_M\left[(u_{rk}\gamma_{dk} - \hat{u}_{(dr)k})^2\right] \leq \sum_{k\in U}(1/\pi_k - 1)\, E_M\left[(u_{rk}\gamma_{dk} - 0)^2\right],$$

where $E_M\left[(u_{rk}\gamma_{dk} - 0)^2\right] = \sigma^2_{rk}\gamma_{dk}$. Replacing the terms $E_M\left[(u_{rk}\gamma_{dk} - \hat{u}_{(dr)k})^2\right]$ with $\sigma^2_{rk}\gamma_{dk}$ in expression (A1.5), the AAV (4.3) is inflated. The approximation $\hat{u}_{(dr)k} = 0$ implies that $b_{(dr)k}(\boldsymbol{\pi}) = c_{(dr)k}(\boldsymbol{\pi}) = 0$. Finally, we emphasize that in most cases, the upward is slight, since the $\hat{u}_{(dr)k}$ are obtained by the $\mathbf{z}_k$ variables that generally have a very low predictive power for the $u_{rk}\gamma_{dk}$ values (see Section 4). In these situations $\hat{u}_{(dr)k} \cong (1/N)\sum_{k\in U} u_{rk}\gamma_{dk} \cong 0$. So $E_M(u_{rk}\gamma_{dk}\hat{u}_{(dr)k}) \cong 0$ and $E_M(\hat{u}_{(dr)k})^2 \cong 0$.

Appendix A4

Proof of expression (4.7)

In this case, each $\boldsymbol{\delta}_k$ vector has $H - 1$ zero elements and 1 element equal to 1 (corresponding to the planned population to which the unit k belongs). Given the input values, the optimization procedure $\pi_k = \pi_h$ for $k \in U_h$. Under the above assumption, $[\mathbf{A}(\boldsymbol{\pi})]^{-1}$ is a diagonal matrix with the hh^{th} element given by $[\mathbf{A}_{hh}(\boldsymbol{\pi})]^{-1} = \left[N_h\pi_h^2(1/\pi_h - 1)\right]^{-1}$. Considering $\tilde{y}_{rk} = \bar{Y}_{rh}$, expressions (A1.2) and (A1.3) can be reformulated as, respectively,

$$\hat{\bar{y}}_{(dr)k} = \pi_h\boldsymbol{\delta}_k'[\mathbf{A}(\boldsymbol{\pi})]^{-1} N_h\pi_h(1/\pi_h - 1)\bar{Y}_{rh} = \bar{Y}_{rh}. \tag{A4.1}$$

$$\hat{u}_{(dr)k} = \pi_h\boldsymbol{\delta}_k'[\mathbf{A}(\boldsymbol{\pi})]^{-1}\pi_h(1/\pi_h - 1)\sum_{j\in U}u_{rj} = (\pi_h N_h)^{-1}\sum_{j\in U_h}u_{rj}, \tag{A4.2}$$

but $\sum_{j\in U_h}u_{rj} = 0$ as the sum of the residual of a regression model.

Using the formulae (A4.1) and (A4.2), expression (4.5) is given by

$$\begin{aligned}\text{AAV}(\hat{t}_{(dr)}) &= [N/(N-H)]\sum_h\left(\frac{1}{\pi_h} - 1\right)\sum_{k\in U_h}E_M(u_{rk}\gamma_{dk})^2 \\ &= [N/(N-H)]\sum_{d=1}^{D}\sum_{h\in H_d}\sigma^2_{rh}N_h(N_h/n_h - 1),\end{aligned}$$

since $\pi_h = n_h/N_h$ and expression (4.7) may be obtained.

References

Bethel, J. (1989). Sample allocation in multivariate surveys. *Survey Methodology*, 15, 1, 47-57.

Boyd, S., and Vanderberg, L. (2004). *Convex Optimization*. Cambridge University Press.

Breidt, F.J., and Chauvet, G. (2011). Improved variance estimation for balanced samples drawn via the cube method. *Journal of Statistical Planning and Inference*, 141, 479-487.

Chauvet, G., Bonnéry, D. and Deville, J.-C. (2011). Optimal inclusion probabilities for balanced sampling. *Journal of Statistical Planning and Inference*, 141, 984-994.

Choudhry, G.H., Rao, J.N.K. and Hidiroglou, M.A. (2012). On sample allocation for efficient domain estimation. *Survey Methodology*, 18, 1, 23-29.

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*. New York: John Wiley & Sons, Inc.

Deville, J.-C., and Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, 91, 893-912.

Deville, J.-C., and Tillé, Y. (2005). Variance approximation under balanced sampling, *Journal of Statistical Planning and Inference*, 128, 569-591.

Dykstra R. and Wollan P. (1987). Finding I-projections subject to a finite set of linear inequality constraints, *Applied Statistics*, 36, 377-383.

Ernst, L.R. (1989). Further applications of linear programming to sampling problems. *Proceedings of the Survey Research Methods Section,* American Statistical Association, 625-631.

Falorsi, P.D., and Righi, P. (2008). A balanced sampling approach for multi-way stratification designs for small area estimation. *Survey Methodology*, 34, 2, 223-234.

Falorsi, P.D., Orsini, D. and Righi, P. (2006). Balanced and coordinated sampling designs for small domain estimation. *Statistics in Transition*, 7, 1173-1198.

Gonzalez, J.M., and Eltinge, J.L. (2010). Optimal survey design: A review. *Section on Survey Research Methods – JSM 2010*, October.

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

Khan, M.G.M., Mati, T. and Ahsan, M.J. (2010). An optimal multivariate stratified sampling design using auxiliary information: An integer solution using goal programming approach. *Journal of Official Statistics*, 26, 695-708.

Kokan, A., and Khan, S. (1967). Optimum allocation in multivariate surveys: An analytical solution. *Journal of the Royal Statistical Society, Series B*, 29, 115-125.

Lu, W., and Sitter, R.R. (2002). Multi-way stratification by linear programming made practical. *Survey Methodology*, 28, 2, 199-207.

Nedyalkova, D., and Tillé, Y. (2008). Optimal sampling and estimation strategies under the linear model. *Biometrika*, 95, 521-537.

Tillé, Y. (2006). *Sampling Algorithms*. Springer-Verlag, New York.

Tillé, Y., and Favre, A.-C. (2005). Optimal allocation in balanced sampling. *Statistics and Probability Letters*, 74, 31-37.

Winkler, W.E. (2001). Multi-way survey stratification and sampling. *Research Report Series*, Statistics #2001-01. Statistical Research Division U.S. Bureau of the Census Washington D.C. 20233.

Survey Methodology, June 2015
Vol. 41, No. 1, pp. 237-262
Statistics Canada, Catalogue No. 12-001-X

An efficient estimation method for matrix survey sampling

Takis Merkouris[1]

Abstract

Matrix sampling, often referred to as split-questionnaire, is a sampling design that involves dividing a questionnaire into subsets of questions, possibly overlapping, and then administering each subset to one or more different random subsamples of an initial sample. This increasingly appealing design addresses concerns related to data collection costs, respondent burden and data quality, but reduces the number of sample units that are asked each question. A broadened concept of matrix design includes the integration of samples from separate surveys for the benefit of streamlined survey operations and consistency of outputs. For matrix survey sampling with overlapping subsets of questions, we propose an efficient estimation method that exploits correlations among items surveyed in the various subsamples in order to improve the precision of the survey estimates. The proposed method, based on the principle of best linear unbiased estimation, generates composite optimal regression estimators of population totals using a suitable calibration scheme for the sampling weights of the full sample. A variant of this calibration scheme, of more general use, produces composite generalized regression estimators that are also computationally very efficient.

Key Words: Best linear unbiased estimator; Calibration; Composite estimator; Generalized regression estimator; Non-nested matrix sampling; Split-questionnaire.

1 Introduction

Matrix sampling is a sampling design in which a long questionnaire is divided into subsets of questions (items), possibly overlapping, and each subset is then administered to one or more distinct random subsamples of an initial sample. In its various forms this design may serve a variety of purposes, such as reducing the length and cost of the survey process and addressing concerns related to respondent burden and data quality associated with a long questionnaire. Matrix sampling has been applied or explored in various fields, primarily in educational assessment and public health studies. A review of previous research on matrix sampling, with discussion of the issues arising in its implementation in surveys, is given in Gonzalez and Eltinge (2007). For recent work on design and estimation for matrix survey sampling, motivated by the potential benefits of such sampling schemes in large scale surveys, see Raghunathan and Grizzle (1995), Thomas, Raghunathan, Schenker, Katzoff and Johnson (2006), Gonzalez and Eltinge (2008), Chipperfield and Steel (2009, 2011), and references therein. Among the many matrix sampling designs explored in the literature, we distinguish the following four principal designs varying in the number of subsamples and the number of sub-questionnaires (overlapping or not) administered to each subsample.

(a) Different (non-overlapping) sets of questions are administered to different subsamples.

(b) An additional core set of questions is administered to all subsamples in design (a). There are several reasons for including a core set of items in all subsamples: High precision may be required for some items of special interest; some other items (e.g., demographic characteristics) define subpopulations and may be used in cross-tabulations of survey results; the correlation of the core items with the rest of items may be used to enhance the precision of estimates for all items.

1. Takis Merkouris, Department of Statistics, Athens University of Economics and Business, Patision 76, Athens 10434, Greece. E-mail: merkouris@aueb.gr.

(c) A variant of design (a) involving an additional subsample that receives the full questionnaire. It may be viewed as a generalization of two-phase sampling design. The motivation for this design is to allow for analysis of interaction between sets of questions, by having responses to all questions from the units of the additional sample, and to enable more efficient estimation.

(d) An extension of design (c), in which the core set of questions is administered to all subsamples. It embodies all features of the previous three designs.

A current trend in survey planning relates to a variant of matrix sampling in which a number of distinct surveys with overlapping content are integrated for the benefit of streamlined survey operations, harmonized survey content and data consistency, as well as improved estimation. In this nonstandard matrix sampling framework, the distinct surveys may use subsamples of a large master sample or independent samples from the same population. Such sampling schemes are actively being researched or implemented in various statistical agencies; see, for example, the integration of household surveys in the British Office of National Statistics (Smith 2009) and in the Australian Bureau of Statistics (2011). Although such integration may be viewed as a reverse process to splitting a questionnaire, the structure of the design with respect to the collection of different subsets of data items from different samples is essentially the same as in the standard framework. In the particular case where the samples from constituent surveys are independent, possibly with different sampling designs, the designs (b), (c) and (d) could be characterized as non-nested matrix sampling designs. It is to be noted that the advantages of matrix sampling are not always contingent on using subsamples (necessarily dependent) of an initial sample. It may be more practical in certain situations to use independent samples, notwithstanding the possibility of a negligible sample overlap.

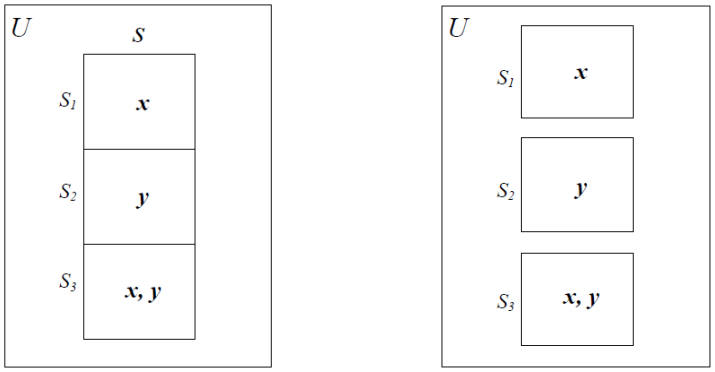
In this paper we address the estimation problem in matrix sampling, namely the loss of precision of survey estimates due to not collecting all data items from all sample units. In the nonstandard matrix sampling of the preceding paragraph, the estimation problem is the improvement of the precision of estimates for each constituent survey. For matrix sampling designs (b), (c) and (d), involving overlapping subsets of questions, a dual estimation task is to combine data on common items from different subsamples for improved estimation, and to exploit correlations among items surveyed in different subsamples for more efficient estimation for all items. To this aim, estimation involving imputation of the missing values caused by the omitted items in each subquestionnaire has been explored in Raghunathan and Grizzle (1995) and Thomas et al. (2006). Estimation using a simple weighting adjustment that combines data on common items has been considered by Gonzalez and Eltinge (2008). In the particular case of non-nested design (b), the estimation problem of combining data from independent samples has also been dealt with in the literature; see, for example, Renssen and Nieuwenbroek (1997), Houbiers (2004), Merkouris (2004, 2010), Wu (2004) and Kim and Rao (2012). Non-nested design (d) has been considered in Renssen (1998). We propose an efficient estimation method, based on the principle of best linear unbiased estimation, which produces composite optimal regression estimators of totals by means of a suitable calibration procedure for the sampling weights of the combined sample, when the second-order sample inclusion probabilities are known. A variant of this calibration procedure of more general applicability produces composite generalized regression estimators, which for certain sampling settings are optimal regression estimators. The method exploits correlations of items across the subsamples to improve the efficiency of estimators even for items surveyed in all subsamples. It is also operationally

very convenient, producing estimates for all items at population or domain level by means of a simple adaptation of the standard calibration system commonly used in statistical agencies. Introducing here the method, we study in detail the principal designs (c) and (d). Adaptations to more general designs are fairly straightforward.

In the following Section 2 and Section 3 we describe the proposed method for design (c). The application of the method to design (d) is described in Section 4. Domain estimation is dealt with in Section 5. A simulation study is presented in Section 6. We conclude with a discussion in Section 7.

2 Composite optimal regression estimation for design (c)

A general estimation method for matrix sampling is illustrated for design (c) through the simplest setting involving three samples S_1, S_2 and S_3 with arbitrary designs and sizes n_1, n_2, n_3, which may be subsamples of an initial sample of size $n = n_1 + n_2 + n_3$ from a population labeled $U = 1, \ldots, k, \ldots, N$, or may be drawn independently from U. A $p-$dimensional vector of variables $\mathbf{x}$ and a $q-$dimensional vector of variables $\mathbf{y}$ are surveyed in S_1 and S_2, respectively, and both vectors are surveyed in S_3. These two modes of matrix sampling, depicted in Figure 2.1, will henceforth be referred to as nested and non-nested matrix sampling, respectively, in analogy with the nested and non-nested two-phase sampling (Hidiroglou 2001).



Nested design Non-nested design

Figure 2.1 Nested and non-nested matrix sampling design (c)

We denote by $\mathbf{w}_i$ the vector of design weights for sample $S_i, i = 1, 2, 3$, and by $\mathbf{X}_i$ and $\mathbf{Y}_i$ the sample matrices of $\mathbf{x}$ and $\mathbf{y}$, the subscripts indicating the sample. We obtain simple Horvitz-Thompson (HT) estimators $\hat{\mathbf{X}}_1 (= \mathbf{X}_1' \mathbf{w}_1)$ and $\hat{\mathbf{X}}_3$ of the population total $\mathbf{t}_\mathbf{x}$ of $\mathbf{x}$, using S_1 and S_3, respectively, and HT estimators $\hat{\mathbf{Y}}_2$ and $\hat{\mathbf{Y}}_3$ of the total $\mathbf{t}_\mathbf{y}$ of $\mathbf{y}$, using S_2 and S_3. For more efficient estimation of the totals $\mathbf{t}_\mathbf{x}$ and $\mathbf{t}_\mathbf{y}$ we seek composite estimators that combine all the available information on $\mathbf{x}$ and $\mathbf{y}$ in the

three samples. Such composite estimators that are best linear unbiased estimators (BLUE), i.e., minimum-variance linear unbiased combinations of the four estimators $\hat{\mathbf{X}}_1, \hat{\mathbf{Y}}_2, \hat{\mathbf{X}}_3$ and $\hat{\mathbf{Y}}_3$, are denoted by $\hat{\mathbf{X}}^B$ and $\hat{\mathbf{Y}}^B$ and given in matrix form by

$$\begin{pmatrix} \hat{\mathbf{X}}^B \\ \hat{\mathbf{Y}}^B \end{pmatrix} = \boldsymbol{\mathcal{P}} \begin{pmatrix} \hat{\mathbf{X}}_1 \\ \hat{\mathbf{Y}}_2 \\ \hat{\mathbf{X}}_3 \\ \hat{\mathbf{Y}}_3 \end{pmatrix}, \tag{2.1}$$

where $\boldsymbol{\mathcal{P}} = (\mathbf{W}'\mathbf{V}^{-1}\mathbf{W})^{-1}\mathbf{W}'\mathbf{V}^{-1}$, the matrix $\mathbf{W}$ satisfies $E\left[(\hat{\mathbf{X}}_1', \hat{\mathbf{Y}}_2', \hat{\mathbf{X}}_3', \hat{\mathbf{Y}}_3')'\right] = \mathbf{W}(\mathbf{t}_\mathbf{x}', \mathbf{t}_\mathbf{y}')'$ and has entries 1's and 0's, and $\mathbf{V}$ is the variance-covariance matrix of $(\hat{\mathbf{X}}_1', \hat{\mathbf{Y}}_2', \hat{\mathbf{X}}_3', \hat{\mathbf{Y}}_3')'$. This estimation method was proposed by Chipperfield and Steel (2009), who provided analytical expressions of the BLUE for scalars x and y in non-nested matrix sampling, assuming simple random sampling and known $\mathbf{V}$. Such an approach to composite estimation has been explored also in a different context of survey sampling; see Wolter (1979), Jones (1980) and Fuller (1990). In general, computation of the BLUE given by (2.1) is not at all practical, as the computation of an estimated matrix $\mathbf{V}$ (and its inverse) in $\boldsymbol{\mathcal{P}}$ would be quite laborious, especially if the number of variables or the sizes of the samples were large; it would be prohibitive if estimates for subpopulations were also required. Of course, the problem would become more difficult with more samples involved.

A more practical formulation of this estimation procedure is as follows. First, we express the composite estimators in (2.1) explicitly as linear combinations of the HT estimators $\hat{\mathbf{X}}_1, \hat{\mathbf{Y}}_2, \hat{\mathbf{X}}_3$ and $\hat{\mathbf{Y}}_3$, i.e.,

$$\begin{aligned} \hat{\mathbf{X}}^B &= \mathbf{B}_{1\mathbf{x}}\hat{\mathbf{X}}_1 + \mathbf{B}_{2\mathbf{x}}\hat{\mathbf{Y}}_2 + \mathbf{B}_{3\mathbf{x}}\hat{\mathbf{X}}_3 + \mathbf{B}_{4\mathbf{x}}\hat{\mathbf{Y}}_3 \\ \hat{\mathbf{Y}}^B &= \mathbf{B}_{1\mathbf{y}}\hat{\mathbf{X}}_1 + \mathbf{B}_{2\mathbf{y}}\hat{\mathbf{Y}}_2 + \mathbf{B}_{3\mathbf{y}}\hat{\mathbf{X}}_3 + \mathbf{B}_{4\mathbf{y}}\hat{\mathbf{Y}}_3. \end{aligned}$$

The condition of unbiasedness, $E(\hat{\mathbf{X}}^B) = \mathbf{t}_\mathbf{x}$ and $E(\hat{\mathbf{Y}}^B) = \mathbf{t}_\mathbf{y}$, implies that $\mathbf{B}_{3\mathbf{x}} = \mathbf{I} - \mathbf{B}_{1\mathbf{x}}$, $\mathbf{B}_{4\mathbf{x}} = -\mathbf{B}_{2\mathbf{x}}$ and $\mathbf{B}_{4\mathbf{y}} = \mathbf{I} - \mathbf{B}_{2\mathbf{y}}$, $\mathbf{B}_{3\mathbf{y}} = -\mathbf{B}_{1\mathbf{y}}$. Thus, $\boldsymbol{\mathcal{P}}$ and $\mathbf{W}$ can be expressed as

$$\boldsymbol{\mathcal{P}} = \begin{pmatrix} \mathbf{B}_{1\mathbf{x}} & \mathbf{B}_{2\mathbf{x}} & \mathbf{I} - \mathbf{B}_{1\mathbf{x}} & -\mathbf{B}_{2\mathbf{x}} \\ \mathbf{B}_{1\mathbf{y}} & \mathbf{B}_{2\mathbf{y}} & -\mathbf{B}_{1\mathbf{y}} & \mathbf{I} - \mathbf{B}_{2\mathbf{y}} \end{pmatrix}, \quad \mathbf{W}' = \begin{pmatrix} \mathbf{I} & \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} & \mathbf{I} \end{pmatrix},$$

respectively, and the two composite estimators have necessarily the regression form

$$\begin{aligned} \hat{\mathbf{X}}^B &= \hat{\mathbf{X}}_3 + \mathbf{B}_{1\mathbf{x}}(\hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_3) + \mathbf{B}_{2\mathbf{x}}(\hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3) \\ \hat{\mathbf{Y}}^B &= \hat{\mathbf{Y}}_3 + \mathbf{B}_{1\mathbf{y}}(\hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_3) + \mathbf{B}_{2\mathbf{y}}(\hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3). \end{aligned} \tag{2.2}$$

Then writing $\boldsymbol{\mathcal{P}} = (\boldsymbol{\mathcal{B}}, \mathbf{I} - \boldsymbol{\mathcal{B}})$, in obvious notation for matrix $\boldsymbol{\mathcal{B}}$, we can express (2.1) as

$$\begin{pmatrix} \hat{\mathbf{X}}^B \\ \hat{\mathbf{Y}}^B \end{pmatrix} = \boldsymbol{\mathcal{B}} \begin{pmatrix} \hat{\mathbf{X}}_1 \\ \hat{\mathbf{Y}}_2 \end{pmatrix} + (\mathbf{I} - \boldsymbol{\mathcal{B}}) \begin{pmatrix} \hat{\mathbf{X}}_3 \\ \hat{\mathbf{Y}}_3 \end{pmatrix} = \begin{pmatrix} \hat{\mathbf{X}}_3 \\ \hat{\mathbf{Y}}_3 \end{pmatrix} + \boldsymbol{\mathcal{B}} \begin{pmatrix} \hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_3 \\ \hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3 \end{pmatrix}, \tag{2.3}$$

the right-hand side of (2.3) being the matrix form of (2.2). The problem of finding the optimal (variance-minimizing) $\mathcal{P}$ of the BLUE in (2.1) reduces then to that of finding the optimal matrix $\mathcal{B}$ in (2.3). The estimated optimal $\hat{\mathcal{B}}^o$ is given by

$$\hat{\mathcal{B}}^o = -\widehat{\mathrm{Cov}}\left(\begin{pmatrix}\hat{\mathbf{X}}_3\\ \hat{\mathbf{Y}}_3\end{pmatrix}, \begin{pmatrix}\hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_3\\ \hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3\end{pmatrix}\right)\left[\hat{V}\begin{pmatrix}\hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_3\\ \hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3\end{pmatrix}\right]^{-1}, \tag{2.4}$$

and when the three samples are independent it reduces to

$$\hat{\mathcal{B}}^o = \hat{V}\begin{pmatrix}\hat{\mathbf{X}}_3\\ \hat{\mathbf{Y}}_3\end{pmatrix}\left[\hat{V}\begin{pmatrix}\hat{\mathbf{X}}_1\\ \hat{\mathbf{Y}}_2\end{pmatrix} + \hat{V}\begin{pmatrix}\hat{\mathbf{X}}_3\\ \hat{\mathbf{Y}}_3\end{pmatrix}\right]^{-1}. \tag{2.5}$$

In view of (2.3), with such optimal $\hat{\mathcal{B}}^o$ the estimated BLUE in (2.1) involving the estimated $\hat{\mathbf{V}}$, and with $\hat{\mathcal{P}} = (\hat{\mathcal{B}}^o, \mathbf{I} - \hat{\mathcal{B}}^o)$ is a special type of optimal multivariate regression estimator. For the form of the ordinary (single-sample) optimal regression estimator and relevant discussion, see Montanari (1987) and Rao (1994).

Expressing the estimated variance of the HT estimator of a total (see, for example, Särndal, Swensson and Wretman (1992), page 43) as a quadratic form with associated non-negative definite matrix $\mathbf{\Lambda}^0 = \{(\pi_{kl} - \pi_k\pi_l)/\pi_k\pi_l\pi_{kl}\}$, where π_k, π_{kl} are first-and-second order inclusion probabilities, it can be shown after some matrix algebra that

$$\hat{\mathcal{B}}^o = (\mathcal{X}_3'\mathbf{\Lambda}^0\mathcal{X})(\mathcal{X}'\mathbf{\Lambda}^0\mathcal{X})^{-1}, \tag{2.6}$$

where

$$\mathcal{X} = \begin{pmatrix}-\mathbf{X}_1 & \mathbf{0}\\ \mathbf{0} & -\mathbf{Y}_2\\ \mathbf{X}_3 & \mathbf{Y}_3\end{pmatrix} \tag{2.7}$$

is the $n \times (p+q)$ design matrix corresponding to the regression estimator (2.3), $\mathcal{X}_3$ is the matrix $\mathcal{X}$ with the first two rows set equal to zero, and $\mathbf{\Lambda}^0$ is associated with the combined sample $S = S_1 \cup S_2 \cup S_3$, reducing in the non-nested sampling to the block-diagonal matrix $\mathrm{diag}\{\mathbf{\Lambda}_i^0\}$ with $\mathbf{\Lambda}_i^0$ associated with the sample S_i. For the nested design, the probabilities defining $\mathbf{\Lambda}^0$ are products of the probabilities of inclusion in S and the conditional (on S) subsampling probabilities. With this estimated $\hat{\mathcal{B}}^o$, the estimated BLUE in (2.3), called composite optimal regression estimator (COR) and denoted by $\hat{\mathcal{X}}^{\mathrm{COR}}$, is written compactly as $\hat{\mathcal{X}}^{\mathrm{COR}} = \hat{\mathcal{X}}_3 - \hat{\mathcal{B}}^o\hat{\mathcal{X}}[= (\mathcal{X}_3 - \mathcal{X}\hat{\mathcal{B}}^{o'})'\mathbf{w}]$, where $\mathbf{w} = (\mathbf{w}_1', \mathbf{w}_2', \mathbf{w}_3')'$ is the vector of design weights of the combined sample S. It transpires that the COR estimator is in fact the sum of weighted sample regression residuals, and $\hat{\mathcal{B}}^o$ minimizes the quadratic form $(\mathcal{X}_3 - \mathcal{X}\hat{\mathcal{B}}^{o'})'$ $\mathbf{\Lambda}^0(\mathcal{X}_3 - \mathcal{X}\hat{\mathcal{B}}^{o'})$ in these residuals, which is the estimated approximate (large-sample) variance of $\hat{\mathcal{X}}^{\mathrm{COR}}$.

Now, upon writing $\hat{\mathcal{X}}^{\mathrm{COR}}$ as $\hat{\mathcal{X}}^{\mathrm{COR}} = \mathcal{X}_3'[\mathbf{w} + \mathbf{\Lambda}^0\mathcal{X}(\mathcal{X}'\mathbf{\Lambda}^0\mathcal{X})^{-1}(\mathbf{0} - \mathcal{X}'\mathbf{w})]$, it appears that the COR estimator has the form of a calibration estimator (with vector of calibration totals $\mathbf{0} = (\mathbf{0}', \mathbf{0}')'$ of

dimension $(p+q)$), whose components satisfy the constraints $\hat{\mathbf{X}}_1^{\mathrm{COR}} = \hat{\mathbf{X}}_3^{\mathrm{COR}}$ and $\hat{\mathbf{Y}}_2^{\mathrm{COR}} = \hat{\mathbf{Y}}_3^{\mathrm{COR}}$, i.e., calibrated estimates of the same total from two different samples are equal. Indeed, the vector

$$\mathbf{c} = \mathbf{w} + \boldsymbol{\Lambda}^0 \boldsymbol{\mathcal{X}} (\boldsymbol{\mathcal{X}}' \boldsymbol{\Lambda}^0 \boldsymbol{\mathcal{X}})^{-1} (\mathbf{0} - \boldsymbol{\mathcal{X}}' \mathbf{w}), \tag{2.8}$$

is the vector of calibrated weights that minimizes the generalized least-squares distance $(\mathbf{c}-\mathbf{w})' (\boldsymbol{\Lambda}^0)^{-1} (\mathbf{c}-\mathbf{w})$ while satisfying the constraints $\mathbf{X}_1'\mathbf{c}_1 = \mathbf{X}_3'\mathbf{c}_3$ and $\mathbf{Y}_2'\mathbf{c}_2 = \mathbf{Y}_3'\mathbf{c}_3$, where the subcector $\mathbf{c}_i$ corresponds to sample S_i. This follows from a general result for the single-sample case, according to which calibration with the generalized least-squares distance measure may involve an arbitrary $n \times n$ positive definite matrix $\mathbf{R}$ instead of $\boldsymbol{\Lambda}^0$; see Andersson and Thorburn (2005).

We may now write the COR estimator formally as a calibration estimator, $\hat{\boldsymbol{\mathcal{X}}}^{\mathrm{COR}} = \boldsymbol{\mathcal{X}}_3'\mathbf{c}$, and using the subvector of calibrated weights $\mathbf{c}_3$, for sample S_3 only, we obtain the components of $\hat{\boldsymbol{\mathcal{X}}}^{\mathrm{COR}}$ directly in the simple linear forms

$$\hat{\mathbf{X}}^{\mathrm{COR}} = \mathbf{X}_3'\mathbf{c}_3 = \sum\nolimits_{S_3} c_k \mathbf{x}_k; \quad \hat{\mathbf{Y}}^{\mathrm{COR}} = \mathbf{Y}_3'\mathbf{c}_3 = \sum\nolimits_{S_3} c_k \mathbf{y}_k,$$

as in common survey practice. Yet, a decomposition of the vector $\mathbf{c}$ based on the following general lemma on calibration gives an analytic expression of $\hat{\mathbf{X}}^{\mathrm{COR}}$ and $\hat{\mathbf{Y}}^{\mathrm{COR}}$ of the form (2.2), which provides insight into the structure and the efficiency of the COR estimator. The proof of the lemma is given in the Appendix.

Lemma 1 *Let $\boldsymbol{\mathcal{X}}$ be a design matrix of dimension $n \times (p+q)$ and of full rank and written in partition form $(\boldsymbol{\mathcal{X}}, \boldsymbol{\Psi})$, with corresponding vector of calibration totals $\mathbf{t}_{\boldsymbol{\mathcal{X}}} = (\mathbf{t}_{\boldsymbol{\mathcal{X}}}', \mathbf{t}_{\boldsymbol{\Psi}}')'$, and let $\mathbf{R}$ be any positive definite matrix of dimension $n \times n$. Then the vector of calibrated weights $\mathbf{c} = \mathbf{w} + \mathbf{R}\boldsymbol{\mathcal{X}} (\boldsymbol{\mathcal{X}}'\mathbf{R}\boldsymbol{\mathcal{X}})^{-1} (\mathbf{t}_{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}}'\mathbf{w})$, obtained from the calibration procedure involving the distance measure $(\mathbf{c}-\mathbf{w})' \mathbf{R}^{-1} (\mathbf{c}-\mathbf{w})$ and the constraint $\boldsymbol{\mathcal{X}}'\mathbf{c} = \mathbf{t}_{\boldsymbol{\mathcal{X}}}$, can be decomposed as*

$$\mathbf{c} = \mathbf{w} + \mathbf{L}_{\boldsymbol{\Psi}} \boldsymbol{\mathcal{X}} (\boldsymbol{\mathcal{X}}' \mathbf{L}_{\boldsymbol{\Psi}} \boldsymbol{\mathcal{X}})^{-1} [\mathbf{t}_{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}}'\mathbf{w}] + \mathbf{L}_{\boldsymbol{\mathcal{X}}} \boldsymbol{\Psi} (\boldsymbol{\Psi}' \mathbf{L}_{\boldsymbol{\mathcal{X}}} \boldsymbol{\Psi})^{-1} [\mathbf{t}_{\boldsymbol{\Psi}} - \boldsymbol{\Psi}'\mathbf{w}], \tag{2.9}$$

where $\mathbf{L}_{\boldsymbol{\mathcal{X}}} = \mathbf{R}(\mathbf{I} - \mathbf{P}_{\boldsymbol{\mathcal{X}}})$ with $\mathbf{P}_{\boldsymbol{\mathcal{X}}} = \boldsymbol{\mathcal{X}} (\boldsymbol{\mathcal{X}}'\mathbf{R}\boldsymbol{\mathcal{X}})^{-1} \boldsymbol{\mathcal{X}}'\mathbf{R}$, and $\mathbf{L}_{\boldsymbol{\Psi}} = \mathbf{R}(\mathbf{I} - \mathbf{P}_{\boldsymbol{\Psi}})$ with $\mathbf{P}_{\boldsymbol{\Psi}} = \boldsymbol{\Psi} (\boldsymbol{\Psi}'\mathbf{R}\boldsymbol{\Psi})^{-1} \boldsymbol{\Psi}'\mathbf{R}$. The vector $\mathbf{c}$ can be written as

$$\mathbf{c} = \mathbf{c}_{\boldsymbol{\Psi}} + \mathbf{L}_{\boldsymbol{\Psi}} \boldsymbol{\mathcal{X}} (\boldsymbol{\mathcal{X}}' \mathbf{L}_{\boldsymbol{\Psi}} \boldsymbol{\mathcal{X}})^{-1} [\mathbf{t}_{\boldsymbol{\mathcal{X}}} - \boldsymbol{\mathcal{X}}'\mathbf{c}_{\boldsymbol{\Psi}}], \tag{2.10}$$

where the vector

$$\mathbf{c}_{\boldsymbol{\Psi}} = \mathbf{w} + \mathbf{R}\boldsymbol{\Psi} (\boldsymbol{\Psi}'\mathbf{R}\boldsymbol{\Psi})^{-1} [\mathbf{t}_{\boldsymbol{\Psi}} - \boldsymbol{\Psi}'\mathbf{w}]$$

is generated by calibration of the design weights involving only $\boldsymbol{\Psi}$ and $\mathbf{t}_{\boldsymbol{\Psi}}$. By symmetry,

$$\mathbf{c} = \mathbf{c}_{\boldsymbol{\mathcal{X}}} + \mathbf{L}_{\boldsymbol{\mathcal{X}}} \boldsymbol{\Psi} (\boldsymbol{\Psi}' \mathbf{L}_{\boldsymbol{\mathcal{X}}} \boldsymbol{\Psi})^{-1} [\mathbf{t}_{\boldsymbol{\Psi}} - \boldsymbol{\Psi}'\mathbf{c}_{\boldsymbol{\mathcal{X}}}], \tag{2.11}$$

where

$$\mathbf{c}_{\mathcal{X}} = \mathbf{w} + \mathbf{R}\mathcal{X}(\mathcal{X}'\mathbf{R}\mathcal{X})^{-1}[\mathbf{t}_{\mathcal{X}} - \mathcal{X}'\mathbf{w}].$$

Now, if $\mathcal{X}$ is as in (2.7), with corresponding vector of calibration totals $\mathbf{t}_{\mathcal{X}} = (\mathbf{0}', \mathbf{0}')'$, and if $\mathbf{R} = \mathbf{\Lambda}^0$, then it follows from (2.9) that (2.8) can be written in the form

$$\mathbf{c} = \mathbf{w} + \mathbf{L}_{\mathbf{\Psi}}\mathbf{X}(\mathbf{X}'\mathbf{L}_{\mathbf{\Psi}}\mathbf{X})^{-1}\left[\hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_3\right] + \mathbf{L}_{\mathbf{x}}\mathbf{\Psi}(\mathbf{\Psi}'\mathbf{L}_{\mathbf{x}}\mathbf{\Psi})^{-1}\left[\hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3\right],$$

and thus

$$\begin{aligned} \hat{\mathbf{X}}^{\mathrm{COR}} &= \mathbf{X}_3'\mathbf{c}_3 = \hat{\mathbf{X}}_3 + \hat{\mathbf{B}}_{1\mathbf{x}}^{o}\left(\hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_3\right) + \hat{\mathbf{B}}_{2\mathbf{x}}^{o}\left(\hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3\right) \\ &= \hat{\mathbf{B}}_{1\mathbf{x}}^{o}\hat{\mathbf{X}}_1 + \left(\mathbf{I} - \hat{\mathbf{B}}_{1\mathbf{x}}^{o}\right)\hat{\mathbf{X}}_3 + \hat{\mathbf{B}}_{2\mathbf{x}}^{o}\left(\hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3\right), \end{aligned} \tag{2.12}$$

in obvious notation for $\hat{\mathbf{B}}_{1\mathbf{x}}^{o}$ and $\hat{\mathbf{B}}_{2\mathbf{x}}^{o}$. A similar expression is obtained for $\hat{\mathbf{Y}}^{\mathrm{COR}}$. It is seen from (2.12) that the COR estimator $\hat{\mathbf{X}}^{\mathrm{COR}}$ of $\mathbf{t}_{\mathbf{x}}$ is approximately (for large samples) unbiased, and derives its efficiency from combining the two elementary estimators $\hat{\mathbf{X}}_1$ and $\hat{\mathbf{X}}_3$ (pooling information from samples S_1 and S_3) and from borrowing strength from sample S_2 through the correlation between $\mathbf{x}$ and $\mathbf{y}$. In view of (2.10), the estimator $\hat{\mathbf{X}}^{\mathrm{COR}}$ takes the alternative forms

$$\begin{aligned} \hat{\mathbf{X}}^{\mathrm{COR}} &= \mathbf{X}_3'\mathbf{c}_{3\mathbf{\Psi}} + \mathbf{X}_3'\mathbf{L}_{\mathbf{\Psi}}\mathbf{X}(\mathbf{X}'\mathbf{L}_{\mathbf{\Psi}}\mathbf{X})^{-1}\left[\mathbf{X}_1'\mathbf{c}_{1\mathbf{\Psi}} - \mathbf{X}_3'\mathbf{c}_{3\mathbf{\Psi}}\right] \\ &= \hat{\mathbf{X}}_3^{\mathrm{OR}} + \hat{\mathbf{B}}_{1\mathbf{x}}^{o}\left[\hat{\mathbf{X}}_1^{\mathrm{OR}} - \hat{\mathbf{X}}_3^{\mathrm{OR}}\right] \\ &= \hat{\mathbf{B}}_{1\mathbf{x}}^{o}\hat{\mathbf{X}}_1^{\mathrm{OR}} + \left(\mathbf{I} - \hat{\mathbf{B}}_{1\mathbf{x}}^{o}\right)\hat{\mathbf{X}}_3^{\mathrm{OR}}, \end{aligned} \tag{2.13}$$

where $\hat{\mathbf{X}}_i^{\mathrm{OR}} = \hat{\mathbf{X}}_i + \mathbf{X}_i'\mathbf{\Lambda}^0\mathbf{\Psi}(\mathbf{\Psi}'\mathbf{\Lambda}^0\mathbf{\Psi})^{-1}\left(\hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3\right)$ are optimal regression (OR) estimators incorporating the regression effect of the last term in (2.12).

In non-nested matrix sampling, $\mathbf{\Lambda}^0 = \mathrm{diag}\{\mathbf{\Lambda}_i^0\}$, $\hat{\mathbf{X}}_1^{\mathrm{OR}} = \hat{\mathbf{X}}_1$, $\hat{\mathbf{X}}_3^{\mathrm{OR}} = \hat{\mathbf{X}}_3 + \widehat{\mathrm{Cov}}(\hat{\mathbf{X}}_3, \hat{\mathbf{Y}}_3)[\hat{V}(\hat{\mathbf{Y}}_2) + \hat{V}(\hat{\mathbf{Y}}_3)]^{-1}[\hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3]$, having estimated approximate variance $\widehat{\mathrm{AV}}\left(\hat{\mathbf{X}}_3^{\mathrm{OR}}\right) = \hat{V}\left(\hat{\mathbf{X}}_3\right) - \widehat{\mathrm{Cov}}(\hat{\mathbf{X}}_3, \hat{\mathbf{Y}}_3)\left[\hat{V}\left(\hat{\mathbf{Y}}_2\right) + \hat{V}\left(\hat{\mathbf{Y}}_3\right)\right]^{-1}\widehat{\mathrm{Cov}}'\left(\hat{\mathbf{X}}_3, \hat{\mathbf{Y}}_3\right)$, and $\hat{\mathbf{B}}_{1\mathbf{x}}^{o} = \widehat{\mathrm{AV}}\left(\hat{\mathbf{X}}_3^{\mathrm{OR}}\right)\left[\hat{V}\left(\hat{\mathbf{X}}_1\right) + \widehat{\mathrm{AV}}\left(\hat{\mathbf{X}}_3^{\mathrm{OR}}\right)\right]^{-1}$ is the coefficient that minimizes the variance $\widehat{\mathrm{AV}}(\hat{\mathbf{X}}^{\mathrm{COR}})$. From the explicit form $\mathbf{I} - \hat{\mathbf{B}}_{1\mathbf{x}}^{o} = \hat{V}\left(\hat{\mathbf{X}}_1\right)\left[\hat{V}\left(\hat{\mathbf{X}}_1\right) + \hat{V}\left(\hat{\mathbf{X}}_3\right) - \widehat{\mathrm{Cov}}\left(\hat{\mathbf{X}}_3, \hat{\mathbf{Y}}_3\right) \times \left[\hat{V}\left(\hat{\mathbf{Y}}_2\right) + \hat{V}\left(\hat{\mathbf{Y}}_3\right)\right]^{-1}\widehat{\mathrm{Cov}}'\left(\hat{\mathbf{X}}_3, \hat{\mathbf{Y}}_3\right)\right]^{-1}$, it is then clear that the stronger the correlation between $\mathbf{x}$ and $\mathbf{y}$ the larger the $\mathbf{I} - \hat{\mathbf{B}}_{1\mathbf{x}}^{o}$ and more weight is given to the less variable component $\hat{\mathbf{X}}_3^{\mathrm{OR}}$. In this connection, it can be easily shown that $\widehat{\mathrm{AV}}(\hat{\mathbf{X}}^{\mathrm{COR}})$ satisfies

$$\widehat{\mathrm{AV}}(\hat{\mathbf{X}}^{\mathrm{COR}})\left[\hat{V}\left(\hat{\mathbf{X}}_1\right)\right]^{-1} = \hat{\mathbf{B}}_{1\mathbf{x}}^{o} < \mathbf{I}, \quad \widehat{\mathrm{AV}}(\hat{\mathbf{X}}^{\mathrm{COR}})\left[\widehat{\mathrm{AV}}\left(\hat{\mathbf{X}}_3^{\mathrm{OR}}\right)\right]^{-1} = \mathbf{I} - \hat{\mathbf{B}}_{1\mathbf{x}}^{o} < \mathbf{I}.$$

These inequalities hold also for any linear combination of the components of each of the estimators involved. The optimal composite regression estimator $\hat{\mathbf{X}}^{\mathrm{COR}}$ is more efficient than each of its two components $\hat{\mathbf{X}}_1$ and $\hat{\mathbf{X}}_3^{\mathrm{OR}}$ by the shown quantities, with the efficiency depending on the strength of the correlation between $\mathbf{x}$ and $\mathbf{y}$. The estimator $\hat{\mathbf{X}}^{\mathrm{COR}}$ is also more efficient than the estimator

$\tilde{\mathbf{X}}^{\mathrm{COR}} = \tilde{\mathbf{B}}^{o}_{1\mathbf{x}}\hat{\mathbf{X}}_1 + (\mathbf{I} - \tilde{\mathbf{B}}^{o}_{1\mathbf{x}})\hat{\mathbf{X}}_3$, with $\tilde{\mathbf{B}}^{o}_{1\mathbf{x}} = \hat{V}(\hat{\mathbf{X}}_3)[\hat{V}(\hat{\mathbf{X}}_1) + \hat{V}(\hat{\mathbf{X}}_3)]^{-1}$, which does not incorporate the information on $\mathbf{y}$ (does not borrow strength from sample S_2) and has estimated variance $\widehat{\mathrm{AV}}(\tilde{\mathbf{X}}^{\mathrm{COR}}) = \hat{V}(\hat{\mathbf{X}}_1)[\hat{V}(\hat{\mathbf{X}}_1) + \hat{V}(\hat{\mathbf{X}}_3)]^{-1}\hat{V}(\hat{\mathbf{X}}_3)$. Indeed, writing the variance $\widehat{\mathrm{AV}}(\hat{\mathbf{X}}^{\mathrm{COR}}) = \hat{V}(\hat{\mathbf{X}}_1)\hat{\mathbf{B}}^{o}_{1\mathbf{x}}$ as $\widehat{\mathrm{AV}}(\hat{\mathbf{X}}^{\mathrm{COR}}) = \hat{V}(\hat{\mathbf{X}}_1)[\hat{V}(\hat{\mathbf{X}}_1) + \hat{V}(\hat{\mathbf{X}}_3)]^{-1}\hat{V}(\hat{\mathbf{X}}_3)\mathbf{E}$, where $\mathbf{E} = \mathbf{E}_1\mathbf{E}_2$ with $\mathbf{E}_1 = [\mathbf{I} - (\hat{V}(\hat{\mathbf{X}}_3))^{-1}$ $\widehat{\mathrm{Cov}}(\hat{\mathbf{X}}_3, \hat{\mathbf{Y}}_3)[\hat{V}(\hat{\mathbf{Y}}_2) + \hat{V}(\hat{\mathbf{Y}}_3)]^{-1}\widehat{\mathrm{Cov}}'(\hat{\mathbf{X}}_3, \hat{\mathbf{Y}}_3)]$ and $\mathbf{E}_2 = [\mathbf{I} - [\hat{V}(\hat{\mathbf{X}}_1) + \hat{V}(\hat{\mathbf{X}}_3)]^{-1}\widehat{\mathrm{Cov}}(\hat{\mathbf{X}}_3, \hat{\mathbf{Y}}_3)[\hat{V}(\hat{\mathbf{Y}}_2) +$ $\hat{V}(\hat{\mathbf{Y}}_3)]^{-1}\widehat{\mathrm{Cov}}'(\hat{\mathbf{X}}_3, \hat{\mathbf{Y}}_3)]^{-1}$, and noticing that $\mathbf{E} \leq \mathbf{I}$, it follows that

$$\widehat{\mathrm{AV}}(\hat{\mathbf{X}}^{\mathrm{COR}})[\widehat{\mathrm{AV}}(\tilde{\mathbf{X}}^{\mathrm{COR}})]^{-1} = \mathbf{E} \leq \mathbf{I},$$

that is, borrowing strength from S_2 reduces the variance of the composite estimator of $\mathbf{t}_{\mathbf{x}}$ by the factor $\mathbf{E}$, which depends on the strength of the correlation between $\mathbf{x}$ and $\mathbf{y}$. It can be easily verified that for two scalar variables x and y and simple random sampling this result reduces to the analogous analytical result on the efficiency of BLUE given in Chipperfield and Steel (2009, page 231). In this simple case $E = [n_1 + n_3][n_3 + n_2(1-\rho^2)]/[(n_1 + n_3)(n_2 + n_3) - n_1 n_2 \rho^2]$, where ρ is the correlation between x and y. As an illustration, assuming equal sample sizes and correlation $\rho = 0.7$, the efficiency gain is 13.96%.

In nested matrix sampling, the two estimators in (2.13) are $\hat{\mathbf{X}}^{\mathrm{OR}}_i = \hat{\mathbf{X}}_i + \widehat{\mathrm{Cov}}(\hat{\mathbf{X}}_i, \hat{\mathbf{\Psi}})$ $[\hat{V}(\hat{\mathbf{\Psi}})]^{-1}[\hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3]$, and $\hat{\mathbf{B}}^{o}_{1x} = [\widehat{\mathrm{AV}}(\hat{\mathbf{X}}^{\mathrm{OR}}_3) - \widehat{\mathrm{AC}}(\hat{\mathbf{X}}^{\mathrm{OR}}_1, \hat{\mathbf{X}}^{\mathrm{OR}}_3)][\widehat{\mathrm{AV}}(\hat{\mathbf{X}}^{\mathrm{OR}}_1) + \widehat{\mathrm{AV}}(\hat{\mathbf{X}}^{\mathrm{OR}}_3) - 2\widehat{\mathrm{AC}}(\hat{\mathbf{X}}^{\mathrm{OR}}_1,$ $\hat{\mathbf{X}}^{\mathrm{OR}}_3)]^{-1}$, where AC denotes approximate covariance. In this case, in addition to the correlation $\rho_{x3,y3}$ between $\hat{\mathbf{X}}_3$ and $\hat{\mathbf{Y}}_3$ in sample S_3, the efficiency of $\hat{\mathbf{X}}^{\mathrm{COR}}$ depends on the estimators' correlations $\rho_{x1,x3}, \rho_{y2,y3}, \rho_{y2,x3}$ due to the dependence of the subsamples. For univariate x and y and with the simplifying assumption of identical designs for the three subsamples (as in equal splitting of the full sample), we obtain some insight through the simple expressions $\widehat{\mathrm{AV}}(\hat{X}^{\mathrm{COR}}) =$ $V(\hat{X}_3)\left[2(1-\rho^2_{x1,x3})(1-\rho_{y2,y3}) - (\rho_{x3,y3} - \rho_{y2,x3})^2\right] \Big/ \left[4(1-\rho_{x1,x3})(1-\rho_{y2,y3}) - (\rho_{x3,y3} - \rho_{y2,x3})^2\right]$, and $\widehat{\mathrm{AV}}(\tilde{X}^{\mathrm{COR}}) = V(\hat{X}_3)(1+\rho_{x1,x3})/2$. Clearly, the estimator $\tilde{X}^{\mathrm{COR}}$, which ignores information on y, is more efficient than the simple average of single-sample estimators of t_x only when there is negative correlation $\rho_{x1,x3}$. The efficiency of $\hat{X}^{\mathrm{COR}}$ relative to $\tilde{X}^{\mathrm{COR}}$

$$\frac{\widehat{\mathrm{AV}}(\hat{X}^{\mathrm{COR}})}{\widehat{\mathrm{AV}}(\tilde{X}^{\mathrm{COR}})} = \frac{4(1-\rho^2_{x1,x3})(1-\rho_{y2,y3}) - 2(\rho_{x3,y3} - \rho_{y2,x3})^2}{4(1-\rho^2_{x1,x3})(1-\rho_{y2,y3}) - (1+\rho_{x1,x3})(\rho_{x3,y3} - \rho_{y2,x3})^2}$$

depends on the sign and size of $\rho_{x1,x3}$ and the size of $|\rho_{x3,y3} - \rho_{y2,x3}|$.

Although the calibration procedure, with vector of calibrated weights (2.8), substantially facilitates the computation of the composite optimal regression estimator for any total of interest, the matrix $\mathbf{\Lambda}^0$ makes the calculations exceedingly demanding, particularly in nested sampling where the subsamples are dependent and thus $\mathbf{\Lambda}^0$ is not $\mathrm{diag}\{\mathbf{\Lambda}^0_i\}$. Besides, the probabilities π_{kl} are not known for most sampling designs. An alternative composite regression estimator that is computationally very efficient is developed in the next section.

3 Composite generalized regression estimation for design (c)

A computationally very convenient, but generally suboptimal, variant of $\hat{\mathcal{B}}^{o}$ in (2.6) is obtained by replacing the matrix $\mathbf{\Lambda}^{0}$ with the diagonal "weighting matrix" $\mathbf{\Lambda}$ having w_{ik}/q_{ik} as ik^{th} diagonal entry, where $\{w_{ik}\}$ are the design weights of S_i and $\{q_{ik}\}$ are positive constants. This gives the multivariate composite generalized regression (CGR) estimator of $(\mathbf{t}_{\mathbf{x}}', \mathbf{t}_{\mathbf{y}}')'$

$$\begin{pmatrix} \hat{\mathbf{X}}^{\text{CGR}} \\ \hat{\mathbf{Y}}^{\text{CGR}} \end{pmatrix} = \hat{\mathcal{B}} \begin{pmatrix} \hat{\mathbf{X}}_1 \\ \hat{\mathbf{Y}}_2 \end{pmatrix} + (\mathbf{I} - \hat{\mathcal{B}}) \begin{pmatrix} \hat{\mathbf{X}}_3 \\ \hat{\mathbf{Y}}_3 \end{pmatrix} = \begin{pmatrix} \hat{\mathbf{X}}_3 \\ \hat{\mathbf{Y}}_3 \end{pmatrix} + \hat{\mathcal{B}} \begin{pmatrix} \hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_3 \\ \hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3 \end{pmatrix}, \tag{3.1}$$

where $\hat{\mathcal{B}} = (\mathcal{X}_3'\mathbf{\Lambda}\mathcal{X})(\mathcal{X}'\mathbf{\Lambda}\mathcal{X})^{-1}$ is the associated matrix regression coefficient. For an extensive discussion of the generalized regression estimator in a single sample, see Särndal et al. (1992, Chapter 6). The CGR estimator may be compactly written as $\hat{\mathcal{X}}^{\text{CGR}} = \hat{\mathcal{X}}_3 - \hat{\mathcal{B}}\hat{\mathcal{X}}\left[= (\mathcal{X}_3 - \mathcal{X}\hat{\mathcal{B}}')' \mathbf{w}\right]$, i.e., as a sum of weighted sample regression residuals. The coefficient $\hat{\mathcal{B}}$ is optimal in the sense of generalized least squares, i.e., it minimizes the quadratic form $(\mathcal{X}_3 - \mathcal{X}\hat{\mathcal{B}}')' \mathbf{\Lambda} (\mathcal{X}_3 - \mathcal{X}\hat{\mathcal{B}}')$ in these residuals. Similarly to the COR estimator, the CGR estimator too can be obtained in calibration form as $\mathcal{X}_3'\mathbf{c}$, where the vector $\mathbf{c} = \mathbf{w} + \mathbf{\Lambda}\mathcal{X}(\mathcal{X}'\mathbf{\Lambda}\mathcal{X})^{-1}(\mathbf{0} - \mathcal{X}'\mathbf{w})$ minimizes the generalized least-squares distance $(\mathbf{c} - \mathbf{w})' \mathbf{\Lambda}^{-1} (\mathbf{c} - \mathbf{w})$ and satisfies the constraints $\hat{\mathbf{X}}_1^{\text{CGR}} = \hat{\mathbf{X}}_3^{\text{CGR}}$ and $\hat{\mathbf{Y}}_2^{\text{CGR}} = \hat{\mathbf{Y}}_3^{\text{CGR}}$. This extends to the present context the well-known equivalence of generalized regression estimation and calibration estimation (Deville and Särndal 1992) for a single-sample setting. Now using the subvector of calibrated weights $\mathbf{c}_3$, for sample S_3 only, we obtain the composite estimators in (3.1) in the simple linear forms $\hat{\mathbf{X}}^{\text{CGR}} = \mathbf{X}_3'\mathbf{c}_3$ and $\hat{\mathbf{Y}}^{\text{CGR}} = \mathbf{Y}_3'\mathbf{c}_3$. Using Lemma 1 and the diagonal structure of $\mathbf{\Lambda}$, it works out that $\hat{\mathbf{X}}^{\text{CGR}}$ can be written as

$$\hat{\mathbf{X}}^{\text{CGR}} = \hat{\mathbf{B}}_{1\mathbf{x}}\hat{\mathbf{X}}_1 + (\mathbf{I} - \hat{\mathbf{B}}_{1\mathbf{x}})\hat{\mathbf{X}}_3^{\text{GR}}, \tag{3.2}$$

where $\hat{\mathbf{X}}_3^{\text{GR}} = \hat{\mathbf{X}}_3 + \mathbf{X}_3'\mathbf{\Lambda}\mathbf{\Psi}(\mathbf{\Psi}'\mathbf{\Lambda}\mathbf{\Psi})^{-1}(\hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3)$ is the generalized regression (GR) counterpart of $\hat{\mathbf{X}}_3^{\text{OR}}$. The matrix regression coefficient $\hat{\mathbf{B}}_{1\mathbf{x}}$ is written explicitly as $\hat{\mathbf{B}}_{1\mathbf{x}} = \mathbf{X}_3'\mathbf{L}_{\mathbf{\Psi}}\mathbf{X}(\mathbf{X}_1'\mathbf{\Lambda}_1\mathbf{X}_1 + \mathbf{X}_3'\mathbf{L}_{\mathbf{\Psi}}\mathbf{X})^{-1}$, where $\mathbf{X}_3'\mathbf{L}_{\mathbf{\Psi}}\mathbf{X} = \mathbf{X}_3'\mathbf{\Lambda}_3\mathbf{X}_3 - \mathbf{X}_3'\mathbf{\Lambda}_3\mathbf{Y}_3(\mathbf{Y}_2'\mathbf{\Lambda}_2\mathbf{Y}_2 + \mathbf{Y}_3'\mathbf{\Lambda}_3\mathbf{Y}_3)^{-1}\mathbf{Y}_3'\mathbf{\Lambda}_3\mathbf{X}_3$. If $\mathbf{x}$ and $\mathbf{y}$ were uncorrelated, or if information on $\mathbf{y}$ was not used in the estimation of $\mathbf{t}_{\mathbf{x}}$, then it would be $\hat{\mathbf{X}}_3^{\text{GR}} = \hat{\mathbf{X}}_3$ and $\hat{\mathbf{B}}_{1\mathbf{x}} = \mathbf{X}_3'\mathbf{\Lambda}_3\mathbf{X}_3(\mathbf{X}_1'\mathbf{\Lambda}_1\mathbf{X}_1 + \mathbf{X}_3'\mathbf{\Lambda}_3\mathbf{X}_3)^{-1}$. But the GR estimator $\hat{\mathbf{X}}_3^{\text{GR}}$ is generally more efficient than the HT estimator $\hat{\mathbf{X}}_3$, and since $\mathbf{X}_1'\mathbf{\Lambda}_1\mathbf{X}_1 + \mathbf{X}_3'\mathbf{L}_{\mathbf{\Psi}}\mathbf{X} < \mathbf{X}_1'\mathbf{\Lambda}_1\mathbf{X}_1 + \mathbf{X}_3'\mathbf{\Lambda}_3\mathbf{X}_3$ (in the partial ordering of non-negative definite matrices), it is clear that more weight is given to $\hat{\mathbf{X}}_3^{\text{GR}}$ in (3.2), through $\mathbf{I} - \hat{\mathbf{B}}_{1\mathbf{x}} = \mathbf{X}_1'\mathbf{\Lambda}_1\mathbf{X}_1(\mathbf{X}_1'\mathbf{\Lambda}_1\mathbf{X}_1 + \mathbf{X}_3'\mathbf{L}_{\mathbf{\Psi}}\mathbf{X})^{-1}$, than would have been given to the component estimator $\hat{\mathbf{X}}_3$ in the simple composite estimator involving only information on $\mathbf{x}$. This suggests that the CGR estimator in (3.2), incorporating information from sample S_2, is a more efficient estimator. Suggestive of the efficiency of $\hat{\mathbf{X}}^{\text{CGR}}$ is also its alternative expression, obtained using (2.11), $\hat{\mathbf{X}}^{\text{CGR}} = \tilde{\mathbf{X}}^{\text{CGR}} + \mathbf{X}_3'\mathbf{L}_{\mathbf{X}}\mathbf{\Psi}(\mathbf{\Psi}'\mathbf{L}_{\mathbf{X}}\mathbf{\Psi})^{-1}\left[\hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3^{\text{GR}}\right]$, where $\tilde{\mathbf{X}}^{\text{CGR}} = \hat{\mathbf{X}}_3 + \mathbf{X}_3'\mathbf{\Lambda}\mathbf{X}(\mathbf{X}'\mathbf{\Lambda}\mathbf{X})^{-1}(\hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_3) = \tilde{\mathbf{B}}_{1\mathbf{x}}\hat{\mathbf{X}}_1 + (\mathbf{I} - \tilde{\mathbf{B}}_{1\mathbf{x}})\hat{\mathbf{X}}_3$ is the composite regression estimator of $\mathbf{t}_{\mathbf{x}}$ using information on $\mathbf{x}$ from S_1 and S_3.

In general, the computationally simpler CGR estimator $(\hat{\mathbf{X}}^{\mathrm{CGR}}, \hat{\mathbf{Y}}^{\mathrm{CGR}})$, involving the coefficient $\hat{\mathcal{B}}$, is less efficient than the optimal composite regression estimator $(\hat{\mathbf{X}}^{\mathrm{COR}}, \hat{\mathbf{Y}}^{\mathrm{COR}})$ which involves the estimated optimal coefficient $\hat{\mathcal{B}}^{o}$ and has the same asymptotic variance as the BLUE in (2.3); the efficiency loss may be larger in nested matrix sampling, for which the matrix $\mathbf{\Lambda}^{0}$ is not block-diagonal. On the other hand, $(\hat{\mathbf{X}}^{\mathrm{COR}}, \hat{\mathbf{Y}}^{\mathrm{COR}})$ may be unstable in small samples, when there is a small number of degrees of freedom available for the estimation of $\hat{\mathcal{B}}^{o}$, which is particularly so in nested matrix sampling; for a discussion of the relative stability of the optimal versus the generalized regression estimator in the single-sample case see Rao (1994) or Montanari (1998). For certain sampling strategies, described in the following theorem, $\hat{\mathcal{B}} = \hat{\mathcal{B}}^{o}$ and the CGR estimator is the COR estimator, and asymptotically is BLUE; the proof is given in the Appendix.

Theorem 1 *Consider the following sampling strategies.*

Non-nested design

(*a*) *For all three samples* S_1, S_2 *and* S_3 *assume stratified simple random sampling without replacement (STRSRS) with sampling fraction* $f_{ih} = n_{ih}/N_{ih}$ *in stratum* h *of sample* i, $h = 1, \ldots, H_i$ *and* N_{ih} *denoting stratum size, and specify the constants* q_{ik} *in* $\mathbf{\Lambda}_i$ *as* $q_{ik} = (n_{ih} - 1)/N_{ih}(1 - f_{ih})$ *for all units of stratum* h. *Furthermore, assume that within each sample the units are sorted by stratum, and consider the augmented design matrix* $\mathcal{Z} = (\mathcal{X}, \mathbf{D})$ *in (2.7), where* $\mathbf{D}$ *is the block diagonal matrix* $\mathrm{diag}\{\mathbf{D}_1, \mathbf{D}_2, \mathbf{D}_3\}$ *and* $\mathbf{D}_i$ *is the diagonal matrix* $\mathrm{diag}\{\mathbf{1}_{i1}, \ldots, \mathbf{1}_{ih}, \ldots, \mathbf{1}_{iH_i}\}$, *with diagonal element* $\mathbf{1}_{ih}$ *being a vector of ones for all units of stratum* h *in sample* S_i, *and consider the corresponding augmented vector of calibration totals* $\mathbf{t}_{\mathcal{Z}} = (\mathbf{0}', \mathbf{0}', \mathbf{N}_1', \mathbf{N}_2', \mathbf{N}_3')'$, *where* $\mathbf{N}_i$ *is the vector of strata sizes for sample* S_i.

(*b*) *For all three samples* S_1, S_2 *and* S_3 *assume stratified Poisson sampling and specify the constants* q_{ik} *in the entries of* $\mathbf{\Lambda}_i$ *as* $q_{ik} = \pi_{ihk}/(1 - \pi_{ihk})$ *for the units of stratum* h, *where* π_{ihk} *is the inclusion probability of unit* k *in stratum* h *of the* i^{th} *survey.*

Nested design

(*a*') *Assume that an initial stratified simple random sample* S *is split by stratum into three simple random subsamples* S_1, S_2 *and* S_3. *Specify the sampling fractions* f_{ih}, *the constants* q_{ik} *in* $\mathbf{\Lambda}_i$, *the design matrix* $\mathcal{Z} = (\mathcal{X}, \mathbf{D})$ *and the vector of calibration totals* $\mathbf{t}_{\mathcal{Z}}$ *as in part* (*a*).

(*b*') *Assume that an initial stratified Poisson sample* S *is randomly split by stratum into three subsamples* S_1, S_2 *and* S_3, *with unequal inclusion probabilities for the units of each subsample. Specify the constants* q_{ik} *in* $\mathbf{\Lambda}_i$ *as* $q_{ik} = \pi_{ihk}/(1 - \pi_{ihk})$ *for the units of stratum* h, *where* π_{ihk} *is the marginal inclusion probability of unit* k *in stratum* h *of the* i^{th} *subsample.*

Under each of strategies (*a*) *and* (*b*), *the calibration procedure with matrix* $\mathbf{\Lambda}$ *in the least-squares distance measure gives the CGR estimator in (3.1) with* $\hat{\mathcal{B}} = \hat{\mathcal{B}}^{o}$, *implying that the CGR estimator is the COR estimator. For* (*a*') *and* (*b*'), *this holds approximately when the strata sampling fractions are approximately zero.*

Corollary 1 *The result of Theorem 1 holds also for the unstratified versions of all four designs. For simple random sampling without replacement (SRS), in particular, the matrix* $\mathbf{D}$ *reduces to the diagonal matrix* $\text{diag}\{\mathbf{1}_1, \mathbf{1}_2, \mathbf{1}_3\}$ *having as its* i^{th} *diagonal element the* n_i*-dimensional unit vector* $\mathbf{1}_i$*, and the vector of calibration totals is then* $\mathbf{t}_{\mathcal{Z}} = (\mathbf{0}', \mathbf{0}', N, N, N)'$.

Corollary 2 *In non-nested sampling, when the sampling design for each of the three samples is one of the designs in* (a) *and* (b) *or one of their unstratified versions, but not the same for all samples, the result of Theorem 1 holds provided that the matrix* $\mathbf{D}$ *in* $\mathcal{Z}$ *and the vector* $\mathbf{t}_{\mathcal{Z}}$ *are reduced so as to correspond only to the samples for which SRS or STRSRS is used.*

The extended calibration scheme in Theorem $1(a, a')$ includes calibration to the stratum sizes (or to the population size in the SRS version), through the inclusion of an intercept for each stratum in the design matrix $\mathcal{X}$. No additional information is used beyond what is assumed in the sampling design in (a) and (a'), and the form of the resulting CGR estimator remains the same as in (3.1) because the HT estimates of the population and strata sizes are exact. The effect of this extended calibration (with the specified values of q_{ik}) is only to convert the CGR coefficient $\hat{\mathcal{B}}$ to the optimal coefficient $\hat{\mathcal{B}}^o$ and, thus, the CGR estimator to the COR estimator. The practical significance of this conversion lies in carrying out optimal composite regression estimation through the much simpler calibration procedure of generalized regression estimation.

Subsampling as in part (a'), with a priori fixed sample sizes, is a natural procedure in matrix sampling involving splitting a questionnaire. In contrast, in the subsampling scheme of part (b') n_i is the expected sample size of S_i, the actual size being random. Unequal subsampling probabilities may be determined adaptively for increased efficiency; see Gonzalez and Eltinge (2008).

The results of Theorem 1 could extend to other sampling designs, e.g., stratified two-stage simple random sampling in non-nested matrix sampling. However, the required adjustments in the matrices $\mathbf{\Lambda}_i$ would not be easier than using directly the matrices $\mathbf{\Lambda}_i^0$ in the calibration to obtain the optimal composite regression estimator.

For sampling designs other than those assumed in Theorem 1, the value of q_{ik} in the entries of $\mathbf{\Lambda}_i$ should be set to $q_{ik} = \tilde{n}_i/(\tilde{n}_1 + \tilde{n}_2 + \tilde{n}_3)$, where $\tilde{n}_i = n_i/d_i$, d_i denoting design effect, to take into account the differential in effective sample sizes among the three samples. If the same design is used for all samples, then $\tilde{n}_i = n_i$. The justification for this adjustment is based on the argument given in Merkouris (2010) for a similar problem of composite regression estimation.

4 Composite estimation for matrix sampling design (d)

4.1 Core set of variables with known totals

We discuss first a special case of the matrix sampling design (d) in which the variables that are common to the three samples have known totals. In this very realistic sampling setting, all samples collect also information on the same vector of auxiliary variables $\mathbf{z}$ for which the vector of population totals $\mathbf{t}_{\mathbf{z}}$ is known. For illustration we consider again three samples, as in Figure 2.1 (but with $\mathbf{z}$ added in all subsamples). Then, the CGR estimator $\hat{\mathbf{X}}^{\text{CGR}}$ in (3.1) may be augmented with the ordinary regression

terms $\hat{\mathbf{B}}_{3\mathbf{x}}(\mathbf{t}_{\mathbf{z}} - \hat{\mathbf{Z}}_1) + \hat{\mathbf{B}}_{4\mathbf{x}}(\mathbf{t}_{\mathbf{z}} - \hat{\mathbf{Z}}_2) + \hat{\mathbf{B}}_{5\mathbf{x}}(\mathbf{t}_{\mathbf{z}} - \hat{\mathbf{Z}}_3)$, where $\hat{\mathbf{Z}}_i, i = 1,2,3$ is the HT estimator of $\mathbf{t}_{\mathbf{z}}$ based on sample S_i; similarly for $\hat{\mathbf{Y}}^{\mathrm{CGR}}$. This estimator has improved efficiency, as it incorporates additional information, and is generated by a calibration procedure that includes the additional three constraints $\hat{\mathbf{Z}}_i^{\mathrm{CGR}} = \mathbf{t}_{\mathbf{z}}$, and has the design matrix $\mathcal{X}$ in (2.7) augmented with the block-diagonal matrix $\mathbf{Z} = \mathrm{diag}\{\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3\}$. In the simplest case when the sample matrices $\mathbf{Z}_i$ reduce to the unit columns $\mathbf{1}_i$ (with corresponding total the size of the population), the calibration scheme is the one specified in Corollary 1 above. As shown in the proof of the next theorem, an application of Lemma 1 to the present calibration procedure, with partitioned design matrix $(\mathcal{X}, \mathbf{Z}), \mathbf{R} = \mathbf{\Lambda}$ and calibration totals $(\mathbf{0}', \mathbf{0}', \mathbf{t}_{\mathbf{z}}', \mathbf{t}_{\mathbf{z}}', \mathbf{t}_{\mathbf{z}}')'$, gives a modified CGR form of (3.1) with GR estimators incorporating information on $\mathbf{z}$ in place of HT estimators. This is compactly written as $\hat{\mathcal{X}}_3^{\mathrm{GR}} - \hat{\mathcal{B}}\hat{\mathcal{X}}^{\mathrm{GR}}$, where $\hat{\mathcal{X}}_3^{\mathrm{GR}} = \hat{\mathcal{X}}_3 + \mathcal{X}_3'\mathbf{\Lambda}\mathbf{Z}(\mathbf{Z}'\mathbf{\Lambda}\mathbf{Z})^{-1}(\mathbf{t}_{(\mathbf{z})} - \hat{\mathbf{Z}})$, with $\mathbf{t}_{(\mathbf{z})} = (\mathbf{t}_{\mathbf{z}}', \mathbf{t}_{\mathbf{z}}', \mathbf{t}_{\mathbf{z}}')'$, and $\hat{\mathcal{X}}^{\mathrm{GR}}$ expressed similarly, and where $\hat{\mathcal{B}} = [\mathcal{X}_3'\mathbf{\Lambda}(\mathbf{I} - \mathbf{P}_{\mathbf{Z}})\mathcal{X}][\mathcal{X}'\mathbf{\Lambda}(\mathbf{I} - \mathbf{P}_{\mathbf{Z}})\mathcal{X}]^{-1}$ with $\mathbf{P}_{\mathbf{Z}} = \mathbf{Z}(\mathbf{Z}'\mathbf{\Lambda}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{\Lambda}$.

Replacing $\mathbf{\Lambda}$ by $\mathbf{\Lambda}^0$ in the calibration procedure gives the optimal composite regression estimator, compactly written as $\hat{\mathcal{X}}_3^{\mathrm{OR}} - \hat{\mathcal{B}}^o\hat{\mathcal{X}}^{\mathrm{OR}}$, with optimal regression estimators incorporating information on $\mathbf{z}$ in place of GR estimators, and with $\hat{\mathcal{B}}^o = [\mathcal{X}_3'\mathbf{\Lambda}^0(\mathbf{I} - \mathbf{P}_{\mathbf{Z}}^0)\mathcal{X}][\mathcal{X}'\mathbf{\Lambda}^0(\mathbf{I} - \mathbf{P}_{\mathbf{Z}}^0)\mathcal{X}]^{-1}$ where $\mathbf{P}_{\mathbf{Z}}^0 = \mathbf{Z}(\mathbf{Z}'\mathbf{\Lambda}^0\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{\Lambda}^0$. Noticing that $(\mathbf{I} - \mathbf{P}_{\mathbf{Z}}^0)\mathcal{X}_3$ is the matrix of residuals corresponding to $\hat{\mathcal{X}}_3^{\mathrm{OR}}$ and that $\mathcal{X}_3'\mathbf{\Lambda}^0(\mathbf{I} - \mathbf{P}_{\mathbf{Z}}^0)\mathcal{X} = \mathcal{X}_3'(\mathbf{I} - \mathbf{P}_{\mathbf{Z}}^0)'\mathbf{\Lambda}^0(\mathbf{I} - \mathbf{P}_{\mathbf{Z}}^0)\mathcal{X} = \widehat{\mathrm{AC}}(\hat{\mathcal{X}}_3^{\mathrm{OR}}, \hat{\mathcal{X}}^{\mathrm{OR}})$, and similarly for $\widehat{\mathrm{AV}}(\hat{\mathcal{X}}^{\mathrm{OR}})$, it follows that

$$\hat{\mathcal{B}}^o = -\widehat{\mathrm{AC}}\left[\begin{pmatrix}\hat{\mathbf{X}}_3^{\mathrm{OR}}\\ \hat{\mathbf{Y}}_3^{\mathrm{OR}}\end{pmatrix}, \begin{pmatrix}\hat{\mathbf{X}}_1^{\mathrm{OR}} - \hat{\mathbf{X}}_3^{\mathrm{OR}}\\ \hat{\mathbf{Y}}_2^{\mathrm{OR}} - \hat{\mathbf{Y}}_3^{\mathrm{OR}}\end{pmatrix}\right]\left[\widehat{\mathrm{AV}}\begin{pmatrix}\hat{\mathbf{X}}_1^{\mathrm{OR}} - \hat{\mathbf{X}}_3^{\mathrm{OR}}\\ \hat{\mathbf{Y}}_2^{\mathrm{OR}} - \hat{\mathbf{Y}}_3^{\mathrm{OR}}\end{pmatrix}\right]^{-1}, \tag{4.1}$$

in analogy with (2.4), or with (2.5) in non-nested sampling. Thus, $\hat{\mathcal{B}}^o$ is optimal in the sense of minimizing the approximate variance of the estimator $\hat{\mathcal{X}}_3^{\mathrm{OR}} - \hat{\mathcal{B}}^o\hat{\mathcal{X}}^{\mathrm{OR}}$, which is then asymptotically BLUE. An alternative estimator, of weaker optimality, has the form $\hat{\mathcal{X}}_3^{\mathrm{GR}} - \hat{\mathcal{B}}^{wo}\hat{\mathcal{X}}^{\mathrm{GR}}$, where the coefficient $\hat{\mathcal{B}}^{wo} = [\mathcal{X}_3'(\mathbf{I} - \mathbf{P}_{\mathbf{Z}})'\mathbf{\Lambda}^0(\mathbf{I} - \mathbf{P}_{\mathbf{Z}})\mathcal{X}][\mathcal{X}'(\mathbf{I} - \mathbf{P}_{\mathbf{Z}})'\mathbf{\Lambda}^0(\mathbf{I} - \mathbf{P}_{\mathbf{Z}})\mathcal{X}]^{-1}$ has the form (4.1) but with GR estimators in place of OR estimators. This estimator, differing from the CGR only in the regression coefficient, is optimal in the restricted sense of being the composite of GR estimators incorporating information on $\mathbf{z}$ that has minimum approximate variance. In general, this later composite estimator cannot be obtained as a calibration estimator. The following theorem gives conditions under which the CGR estimator is optimal in one of the two senses in non-nested matrix sampling; the proof is given in the Appendix. The nested sampling version of the theorem, with subsampling schemes and proof as in Theorem 1, is omitted for brevity.

Theorem 2 *Consider the following sampling strategies.*

(*a*) *For all three samples S_1, S_2 and S_3 assume SRS with sampling fractions $f_i = n_i/N$, and specify all constants q_{ik} in $\mathbf{\Lambda}_i$ as $q_{ik} = (n_i - 1)/N(1 - f_i)$. Consider the augmented design matrix $\mathcal{Z} = (\mathcal{X}, \mathbf{Z})$ in (2.7), where $\mathbf{Z} = \mathrm{diag}\{\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3\}$, and with the corresponding augmented vector*

of calibration totals $\mathbf{t}_{\mathcal{Z}} = (\mathbf{0}', \mathbf{0}', \mathbf{t}_{\mathbf{z}}', \mathbf{t}_{\mathbf{z}}', \mathbf{t}_{\mathbf{z}}')'$. *Further, suppose that* $\mathbf{Z}_i \mathbf{h}_i = \mathbf{1}$ *for constant vectors* $\mathbf{h}_i$.

Then, the calibration procedure gives the CGR as $\hat{\mathcal{X}}_3^{\mathrm{GR}} - \hat{\mathcal{B}}\hat{\mathcal{X}}^{\mathrm{GR}} = \hat{\mathcal{X}}_3^{\mathrm{GR}} - \hat{\mathcal{B}}^{wo}\hat{\mathcal{X}}^{\mathrm{GR}}$, *i.e., the CGR estimator is the optimal composite of GR estimators incorporating information on* $\mathbf{z}$.

(*b*) *For all three samples* S_1, S_2 *and* S_3 *assume STRSRS with sampling fraction* $f_{ih} = n_{ih}/N_{ih}$ *in stratum* h *of sample* $i, h = 1, \ldots, H_i$ *and* N_{ih} *denoting stratum size, and specify the constants in* $\mathbf{\Lambda}_i$ *as* $q_{ik} = (n_{ih} - 1)/N_h(1 - f_{ih})$ *for all units of stratum* h. *Further, assume that within each sample the units are sorted by stratum, and consider the augmented design matrix* $\mathcal{Z} = (\mathcal{X}, \mathbf{Z}, \mathbf{D})$ *in (2.7), with corresponding augmented vector of calibration totals* $\mathbf{t}_{\mathcal{Z}} = (\mathbf{0}', \mathbf{0}', \mathbf{t}_{\mathbf{z}}', \mathbf{t}_{\mathbf{z}}', \mathbf{t}_{\mathbf{z}}', \mathbf{N}_1', \mathbf{N}_2', \mathbf{N}_3')'$. *The definition of* $\mathbf{D}$ *and* $\mathbf{N}_i$ *is as before.*

Then, the calibration procedure gives the CGR as $\hat{\mathcal{X}}_3^{\mathrm{OR}} - \hat{\mathcal{B}}^{o}\hat{\mathcal{X}}^{\mathrm{OR}}$, *i.e., the CGR estimator is the optimal composite of optimal regression estimators incorporating information on* $\mathbf{z}$.

(*c*) *For all three samples* S_1, S_2 *and* S_3 *assume stratified Poisson sampling and specify the constants* q_{ik} *in the entries of* $\mathbf{\Lambda}_i$ *as* $q_{ik} = \pi_{ihk}/(1 - \pi_{ihk})$ *for the units of stratum* h.

Then, the calibration procedure, with $\mathcal{Z}$ *and* $\mathbf{t}_{\mathcal{Z}}$ *as in* (*a*), *gives the CGR as* $\hat{\mathcal{X}}_3^{\mathrm{GR}} - \hat{\mathcal{B}}\hat{\mathcal{X}}^{\mathrm{GR}} = \hat{\mathcal{X}}_3^{\mathrm{OR}} - \hat{\mathcal{B}}^{o}\hat{\mathcal{X}}^{\mathrm{OR}}$, *i.e., GR and OR estimators are identical, and the CGR estimator is the optimal composite of optimal regression estimators incorporating information on* $\mathbf{z}$.

The condition $\mathbf{Z}_i \mathbf{h}_i = \mathbf{1}$ in (*a*) of Theorem 2 is customarily satisfied when the vector $\mathbf{z}$ contains categorical variables. Results analogous to Corollaries 1 and 2 of the previous section hold also for parts (*b*) and (*c*) of Theorem 2. Here too, for sampling designs other than those assumed in Theorem 2, the value $q_{ik} = \tilde{n}_i/(\tilde{n}_1 + \tilde{n}_2 + \tilde{n}_3)$ in the entries of $\mathbf{\Lambda}$ should be used.

Finally, by analogy to (3.2), and with the appropriate decomposition of the vector of calibrated weights $\mathbf{c}$, the composite estimator $\hat{\mathbf{X}}^{\mathrm{CGR}}$ takes now the form

$$\hat{\mathbf{X}}^{\mathrm{CGR}} = \hat{\mathbf{B}}_{1\mathbf{x}}\hat{\mathbf{X}}_1^{\mathrm{GR}} + (\mathbf{I} - \hat{\mathbf{B}}_{1\mathbf{x}})\hat{\mathbf{X}}_3^{\mathrm{GR}},$$

where $\hat{\mathbf{X}}_1^{\mathrm{GR}}$ and $\hat{\mathbf{X}}_3^{\mathrm{GR}}$ are GR estimators using information on $\mathbf{z}$ from S_1, and on $\mathbf{y}$ and $\mathbf{z}$ from S_2 and S_3, respectively, and $\hat{\mathbf{B}}_{1\mathbf{x}}$ is the corresponding matrix regression coefficient. Similar is the expression for $\hat{\mathbf{Y}}^{\mathrm{CGR}}$. Of course, $\hat{\mathbf{X}}^{\mathrm{CGR}}$ and $\hat{\mathbf{Y}}^{\mathrm{CGR}}$ can be obtained directly through this modified $\mathbf{c}$ in the simple linear forms $\hat{\mathbf{X}}^{\mathrm{CGR}} = \mathbf{X}_3'\mathbf{c}_3$ and $\hat{\mathbf{Y}}^{\mathrm{CGR}} = \mathbf{Y}_3'\mathbf{c}_3$.

4.2 Core set of variables with unknown totals

We turn now to the case of matrix sampling design (d) in which the variables $\mathbf{z}$ that are common to the three samples have unknown totals. Estimation in this setting includes the construction of a composite estimator of the vector of totals $\mathbf{t}_{\mathbf{z}}$. In line with the formulation of Section 2, composite estimators of $\mathbf{t}_{\mathbf{x}}, \mathbf{t}_{\mathbf{y}}$ and $\mathbf{t}_{\mathbf{z}}$ that are best linear unbiased combinations of the HT estimators $\hat{\mathbf{X}}_1, \hat{\mathbf{Z}}_1, \hat{\mathbf{Y}}_2, \hat{\mathbf{Z}}_2, \hat{\mathbf{X}}_3, \hat{\mathbf{Y}}_3, \hat{\mathbf{Z}}_3$ are given by

$$
\begin{aligned}
\hat{\mathbf{X}}^{B} &= \mathbf{B}_{1\mathbf{x}}\hat{\mathbf{X}}_1 + (\mathbf{I}-\mathbf{B}_{1\mathbf{x}})\hat{\mathbf{X}}_3 + \mathbf{B}_{3\mathbf{x}}(\hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3) + \mathbf{B}_{2\mathbf{x}}(\hat{\mathbf{Z}}_1 - \hat{\mathbf{Z}}_3) + \mathbf{B}_{4\mathbf{x}}(\hat{\mathbf{Z}}_2 - \hat{\mathbf{Z}}_3) \\
\hat{\mathbf{Y}}^{B} &= \mathbf{B}_{3\mathbf{y}}\hat{\mathbf{Y}}_2 + (\mathbf{I}-\mathbf{B}_{3\mathbf{y}})\hat{\mathbf{Y}}_3 + \mathbf{B}_{1\mathbf{y}}(\hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_3) + \mathbf{B}_{2\mathbf{y}}(\hat{\mathbf{Z}}_1 - \hat{\mathbf{Z}}_3) + \mathbf{B}_{4\mathbf{y}}(\hat{\mathbf{Z}}_2 - \hat{\mathbf{Z}}_3) \qquad (4.2) \\
\hat{\mathbf{Z}}^{B} &= \mathbf{B}_{2\mathbf{z}}\hat{\mathbf{Z}}_1 + \mathbf{B}_{4\mathbf{z}}\hat{\mathbf{Z}}_2 + (\mathbf{I}-\mathbf{B}_{2\mathbf{z}}-\mathbf{B}_{4\mathbf{z}})\hat{\mathbf{Z}}_3 + \mathbf{B}_{1\mathbf{z}}(\hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_3) + \mathbf{B}_{3\mathbf{z}}(\hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3).
\end{aligned}
$$

The estimators in (4.2) can be written in the matrix regression form

$$
\begin{pmatrix} \hat{\mathbf{X}}^{B} \\ \hat{\mathbf{Y}}^{B} \\ \hat{\mathbf{Z}}^{B} \end{pmatrix} = \begin{pmatrix} \hat{\mathbf{X}}_3 \\ \hat{\mathbf{Y}}_3 \\ \hat{\mathbf{Z}}_3 \end{pmatrix} + \boldsymbol{\mathcal{B}} \begin{pmatrix} \hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_3 \\ \hat{\mathbf{Z}}_1 - \hat{\mathbf{Z}}_3 \\ \hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3 \\ \hat{\mathbf{Z}}_2 - \hat{\mathbf{Z}}_3 \end{pmatrix}, \qquad (4.3)
$$

with the variance-minimizing matrix of coefficients given by $\boldsymbol{\mathcal{B}} = -\mathrm{Cov}(\mathbf{u}_3, \mathbf{u}_{12} - \mathbf{u}_3^{\star})[V(\mathbf{u}_{12} - \mathbf{u}_3^{\star})]^{-1}$, where $\mathbf{u}_3 = (\hat{\mathbf{X}}_3', \hat{\mathbf{Y}}_3', \hat{\mathbf{Z}}_3')'$, $\mathbf{u}_3^{\star} = (\hat{\mathbf{X}}_3', \hat{\mathbf{Z}}_3', \hat{\mathbf{Y}}_3', \hat{\mathbf{Z}}_3')'$, $\mathbf{u}_{12} = (\hat{\mathbf{X}}_1', \hat{\mathbf{Z}}_1', \hat{\mathbf{Y}}_2', \hat{\mathbf{Z}}_2')'$. With estimated covariance and variance matrices we obtain the estimated optimal matrix $\hat{\boldsymbol{\mathcal{B}}}^{o}$, and (4.3) becomes then an optimal multivariate regression estimator. Then, proceeding as in Section 2, it can be shown that

$$
\hat{\boldsymbol{\mathcal{B}}}^{o} = (\boldsymbol{\mathcal{X}}_{3-}'\boldsymbol{\Lambda}^{0}\boldsymbol{\mathcal{X}})(\boldsymbol{\mathcal{X}}'\boldsymbol{\Lambda}^{0}\boldsymbol{\mathcal{X}})^{-1},
$$

where

$$
\boldsymbol{\mathcal{X}} = \begin{pmatrix} -\mathbf{X}_1 & -\mathbf{Z}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & -\mathbf{Y}_2 & -\mathbf{Z}_2 \\ \mathbf{X}_3 & \mathbf{Z}_3 & \mathbf{Y}_3 & \mathbf{Z}_3 \end{pmatrix} \qquad (4.4)
$$

is the design matrix corresponding to the regression estimator (4.3), $\boldsymbol{\mathcal{X}}_{3-}$ is the matrix $\boldsymbol{\mathcal{X}}$ with the second column eliminated and the first two rows set equal to zero, and $\boldsymbol{\Lambda}^{0}$ is as in Section 2.

Replacing the matrix $\boldsymbol{\Lambda}^{0}$ with the weighting matrix $\boldsymbol{\Lambda}$, gives the generalized regression coefficient $\hat{\boldsymbol{\mathcal{B}}} = (\boldsymbol{\mathcal{X}}_{3-}'\boldsymbol{\Lambda}\boldsymbol{\mathcal{X}})(\boldsymbol{\mathcal{X}}'\boldsymbol{\Lambda}\boldsymbol{\mathcal{X}})^{-1}$, and (4.3) becomes the CGR estimator of $(\mathbf{t}_{\mathbf{x}}', \mathbf{t}_{\mathbf{y}}', \mathbf{t}_{\mathbf{z}}')'$

$$
\begin{pmatrix} \hat{\mathbf{X}}^{\mathrm{CGR}} \\ \hat{\mathbf{Y}}^{\mathrm{CGR}} \\ \hat{\mathbf{Z}}^{\mathrm{CGR}} \end{pmatrix} = \begin{pmatrix} \hat{\mathbf{X}}_3 \\ \hat{\mathbf{Y}}_3 \\ \hat{\mathbf{Z}}_3 \end{pmatrix} + \hat{\boldsymbol{\mathcal{B}}} \begin{pmatrix} \hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_3 \\ \hat{\mathbf{Z}}_1 - \hat{\mathbf{Z}}_3 \\ \hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3 \\ \hat{\mathbf{Z}}_2 - \hat{\mathbf{Z}}_3 \end{pmatrix}. \qquad (4.5)
$$

The estimator (4.5) can be conveniently obtained through a calibration procedure that gives a vector of calibrated weights for the combined sample S having the form $\mathbf{c} = \mathbf{w} + \boldsymbol{\Lambda}\boldsymbol{\mathcal{X}}(\boldsymbol{\mathcal{X}}'\boldsymbol{\Lambda}\boldsymbol{\mathcal{X}})^{-1}(\mathbf{0} - \boldsymbol{\mathcal{X}}'\mathbf{w})$, as before, but now satisfying the additional constraint $\hat{\mathbf{Z}}_1^{\mathrm{CGR}} = \hat{\mathbf{Z}}_2^{\mathrm{CGR}} = \hat{\mathbf{Z}}_3^{\mathrm{CGR}}$. Expression (4.5) is then obtained simply as $\boldsymbol{\mathcal{X}}_{3-}'\mathbf{c}$, based on sample S_3.

The explicit expression (4.2), different for the optimal regression and the generalized regression variants only in the form of the linear coefficients, shows that the composite estimators of $\mathbf{t}_{\mathbf{x}}$ and $\mathbf{t}_{\mathbf{y}}$ are more efficient than their counterparts in matrix sampling design (c), equation (2.2), because they incorporate information on the common variables $\mathbf{z}$, assuming non-zero correlation with $\mathbf{x}$ and $\mathbf{y}$. Particularly remarkable is the expression for the composite estimator of $\mathbf{t}_{\mathbf{z}}$: it involves a linear combination of the three HT estimators of $\mathbf{t}_{\mathbf{z}}$ derived from the three samples, plus the two regression

terms implying additional efficiency through the correlation of $\mathbf{z}$ with $\mathbf{x}$ and $\mathbf{y}$. One would expect the additional terms to be zero because an optimal combination of the three estimators should incorporate all information on $\mathbf{z}$ available in the three samples. In general, however, the associated coefficients are not zero. In non-nested sampling, conditions under which these coefficients are zero are given by the following proposition, the proof of which is given in the Appendix. The result should also hold in nested sampling.

Proposition 1 *The coefficients* $\mathbf{B}_{1\mathbf{z}}$ *and* $\mathbf{B}_{3\mathbf{z}}$ *in the estimator* $\hat{\mathbf{Z}}^B$ *in (4.2) are zero only if*

$$\begin{aligned} \left[V(\hat{\mathbf{Z}}_1)\right]^{-1}\mathrm{Cov}(\hat{\mathbf{X}}_1,\hat{\mathbf{Z}}_1) &= \left[V(\hat{\mathbf{Z}}_3)\right]^{-1}\mathrm{Cov}(\hat{\mathbf{X}}_3,\hat{\mathbf{Z}}_3) \\ \left[V(\hat{\mathbf{Z}}_2)\right]^{-1}\mathrm{Cov}(\hat{\mathbf{Y}}_2,\hat{\mathbf{Z}}_2) &= \left[V(\hat{\mathbf{Z}}_3)\right]^{-1}\mathrm{Cov}(\hat{\mathbf{Y}}_3,\hat{\mathbf{Z}}_3). \end{aligned} \tag{4.6}$$

This can happen only if the sampling designs for the three samples are identical, including equal sample sizes, or only if the sampling design across samples is the same design with equal inclusion probability for all units, but not necessarily with the same sample size.

Noticing that the quantities on each side of the equations (4.6) are regression coefficients, according to Proposition 1 the terms of the estimator $\hat{\mathbf{Z}}^B$ incorporating the correlation of $\mathbf{z}$ with $\mathbf{x}$ and $\mathbf{y}$ are zero only if the effect of the regression of $\mathbf{x}$ and $\mathbf{y}$ on $\mathbf{z}$ is identical in samples S_1 and S_3 and in samples S_2 and S_3, respectively. The essence of this finding is that estimation of $\mathbf{t}_\mathbf{z}$ using only information on $\mathbf{z}$ from the three samples, but ignoring information on $\mathbf{x}$ and $\mathbf{y}$, will be suboptimal when there is differential regression effect of $\mathbf{x}$ and $\mathbf{y}$ on $\mathbf{z}$ in the various samples. The efficiency of $\hat{\mathbf{Z}}^B$ relative to the composite estimator $\tilde{\mathbf{Z}}^B$ that uses only information on $\mathbf{z}$ was possible to gauge in the simple setting involving scalar x, y and z, simple random sampling for S_1 and S_3 and Bernoulli sampling for S_2, and equal sampling rates for all three samples. Then only the first equation of (4.6) holds. After much tedious algebra the efficiency of $\hat{\mathbf{Z}}^B$ relative to $\tilde{\mathbf{Z}}^B$ was derived to be $\left[V(\tilde{\mathbf{Z}}^B) - V(\hat{\mathbf{Z}}^B)/V(\tilde{\mathbf{Z}}^B)\right] = G/H$, with

$$\begin{aligned} G &= 2\left(r_{xz}^2-1\right)\left(r_{yz}cv_y - cv_z\right)^2 \\ H &= \left(cv_z^2+1\right)\left(\left(12-9r_{yz}^2\right)r_{xz}^2 - 3r_{xy}\left(2r_{yz}r_{xz}-1\right)+12\left(r_{yz}^2-1\right)\right)cv_z^2cv_y^2 \\ &+ 2\left(r_{xy}^2+r_{yz}^2\right)cv_y^2 + 8\left(r_{xz}^2-1\right)cv_y^2 - 4r_{yz}r_{xy}r_{xz}cv_y^2 \\ &+ 6\left(r_{xz}^2-1\right)cv_z\left(cv_z - 2r_{yz}cv_y\right) \end{aligned}$$

where r_{xy}, r_{xz} and r_{yz} denote population correlation coefficients, and cv_y, cv_z denote coefficients of variation. Although in this setting the departure from the conditions of Proposition 1 is minimal, different configurations of admissible values for $r_{xy}, r_{xz}, r_{yz}, cv_y$ and cv_z show that the efficiency gain may be substantial, making up for the inefficiency of the HT estimator of $\mathbf{t}_\mathbf{z}$ based on the Bernoulli sample S_2. For example, when $r_{xy} = 0.3, r_{xz} = 0.3, r_{yz} = 0.3$ and $cv_y = 0.1, cv_z = 0.6$, the efficiency gain is 23%. In the case of the composite optimal regression estimator $\hat{\mathbf{Z}}^{\mathrm{COR}}$, with estimated coefficients $\hat{\mathbf{B}}_{1\mathbf{z}}^o$ and $\hat{\mathbf{B}}_{3\mathbf{z}}^o$, the regression coefficients in (4.6) are estimated, and thus the equalities in (4.6) would never hold exactly because of the sample differences. Likewise in the case of the CGR estimator $\hat{\mathbf{Z}}^{\mathrm{CGR}}$, for which equations formally identical to (4.6) are given in terms of sample generalized regression coefficients.

Regarding the efficiency of the CGR estimator (4.5), an exact analogue of Theorem 1 holds in the present setting, with the same sampling strategies for which the CGR estimator is optimal regression estimator and asymptotically BLUE.

Composite estimation for a matrix sampling scheme involving a core set of variables with both known and unknown totals can be carried out using the obvious extended calibration scheme.

5 Domain estimation

Composite estimators for domains (subpopulations) of interest may be readily obtained using the calibrated weights derived in the previous sections, that is, by summing the weighted values of a variable over any domain $U_d \subset U$. For instance, letting $\mathbf{X}_{id}$ denote the matrix $\mathbf{X}_i$, for sample S_i, with the entries of the k^{th} row set equal to 0 if $k \notin U_d$, the CGR estimator of the domain total $\mathbf{t}_{\mathbf{x}d}$ based on the weights of S_3 calibrated with the scheme of design (c) (see Section 3) is given by

$$\hat{\mathbf{X}}_{3d}^{\text{CGR}} = \mathbf{X}_{3d}'\mathbf{c}_3 = \hat{\mathbf{X}}_{3d}^{\text{GR}} + \mathbf{X}_{3d}'\mathbf{L}_{\boldsymbol{\Psi}}\mathbf{X}(\mathbf{X}'\mathbf{L}_{\boldsymbol{\Psi}}\mathbf{X})^{-1}\left[\hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_3^{\text{GR}}\right],$$

where $\hat{\mathbf{X}}_{3d}^{\text{GR}} = \hat{\mathbf{X}}_{3d} + \mathbf{X}_{3d}'\boldsymbol{\Lambda}\boldsymbol{\Psi}(\boldsymbol{\Psi}'\boldsymbol{\Lambda}\boldsymbol{\Psi})^{-1}\left(\hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3\right)$ and the subscript d indicates domain. The CGR estimator $\hat{\mathbf{X}}_{1d}^{\text{CGR}}$ based on sample S_1 is obtained in the same manner. However, unlike the population-level estimator (3.2), resulting from calibration of two estimators to each other at population level, the estimators $\hat{\mathbf{X}}_{1d}^{\text{CGR}}$ and $\hat{\mathbf{X}}_{3d}^{\text{CGR}}$ are not constructed as composites of two domain estimators, based on samples S_1 and S_3, and they are not identical. Moreover, although both $\hat{\mathbf{X}}_{1d}^{\text{CGR}}$ and $\hat{\mathbf{X}}_{3d}^{\text{CGR}}$ incorporate information on $\mathbf{x}$ from samples S_1 and S_3, their construction (non-customized at domain level) may entail some loss of efficiency.

A simple modification of the calibration procedure that leads to efficient composite estimation for all totals of interest involves the augmentation of the design matrix with columns defined at each domain level for the relevant variables. Thus, for design (c) estimation of the domain total $\mathbf{t}_{\mathbf{x}d}$ involves the augmentation of the design matrix $\mathcal{X}$ in (2.7) with the column $(-\mathbf{X}_{1d}', \mathbf{0}', \mathbf{X}_{3d}')'$. The resulting estimator, $\tilde{\mathbf{X}}_d^{\text{CGR}}$, may be written in the forms

$$\begin{aligned}\tilde{\mathbf{X}}_d^{\text{CGR}} &= \hat{\mathbf{X}}_{3d} + \hat{\mathbf{B}}_{1\mathbf{x}d}\left(\hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_3\right) + \hat{\mathbf{B}}_{2\mathbf{x}d}\left(\hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3\right) + \hat{\mathbf{B}}_{3\mathbf{x}d}\left(\hat{\mathbf{X}}_{1d} - \hat{\mathbf{X}}_{3d}\right) \\ &= \hat{\mathbf{B}}_{1\mathbf{x}d}\tilde{\mathbf{X}}_{1d}^{\text{GR}} + \left(\mathbf{I} - \hat{\mathbf{B}}_{1\mathbf{x}d}\right)\tilde{\mathbf{X}}_{3d}^{\text{GR}},\end{aligned} \tag{5.1}$$

where $\tilde{\mathbf{X}}_{1d}^{\text{GR}}$ and $\tilde{\mathbf{X}}_{3d}^{\text{GR}}$ are now the GR domain estimators incorporating the regression effect of the second and third terms of (5.1). Adding another term in (5.1) involving the difference $\hat{\mathbf{Y}}_{2d} - \hat{\mathbf{Y}}_{3d}$ may not improve appreciably the efficiency of $\tilde{\mathbf{X}}_d^{\text{CGR}}$ but will be necessary if estimation of the domain total $\mathbf{t}_{\mathbf{y}d}$ is also required. In any particular situation, the augmentation of the design matrix $\mathcal{X}$ involves only those components of $\mathbf{x}$ or $\mathbf{y}$ for which domain estimates are needed. A possible drawback of this procedure is the additional computational burden, which increases with the number of domains and the variables for which domain estimation is required.

An alternative approach that may be more appropriate when the domain estimates of interest are numerous, involves the separate production of the domain estimates by carrying out the composite calibration only at the domain level. For the domain total $\mathbf{t}_{\mathbf{x}d}$, this would give the domain CGR estimator, in analogy with the population CGR estimator (3.2),

$$\breve{\mathbf{X}}_d^{\mathrm{CGR}} = \breve{\mathbf{B}}_{1\mathbf{x}d}\hat{\mathbf{X}}_{1d} + \left(\mathbf{I} - \breve{\mathbf{B}}_{1\mathbf{x}d}\right)\breve{\mathbf{X}}_{3d}^{\mathrm{GR}},$$

where $\breve{\mathbf{B}}_{1\mathbf{x}d} = \mathbf{X}'_{3d}\mathbf{L}_{\boldsymbol{\Psi}_d}\mathbf{X}_d\left(\mathbf{X}'_d\mathbf{L}_{\boldsymbol{\Psi}_d}\mathbf{X}\right)_d^{-1}$ and $\breve{\mathbf{X}}_{3d}^{\mathrm{GR}} = \hat{\mathbf{X}}_{3d} + \mathbf{X}'_{3d}\boldsymbol{\Lambda}\boldsymbol{\Psi}_d\left(\boldsymbol{\Psi}'_d\boldsymbol{\Lambda}\boldsymbol{\Psi}_d\right)^{-1}\left(\hat{\mathbf{Y}}_{2d} - \hat{\mathbf{Y}}_{3d}\right)$. The efficiency of the joint estimator $\left(\breve{\mathbf{X}}_d^{\mathrm{CGR}}, \breve{\mathbf{Y}}_d^{\mathrm{CGR}}\right)$ over the estimator $\left(\hat{\mathbf{X}}_d^{\mathrm{CGR}}, \hat{\mathbf{Y}}_d^{\mathrm{CGR}}\right)$ can be verified under the conditions of the following proposition (its proof in the Appendix).

Proposition 2 *Under the sampling schemes of Theorem 1,*

$$\widehat{\mathrm{AV}}\begin{pmatrix}\breve{\mathbf{X}}_{3d}^{\mathrm{CGR}}\\ \breve{\mathbf{Y}}_{3d}^{\mathrm{GGR}}\end{pmatrix} < \widehat{\mathrm{AV}}\begin{pmatrix}\hat{\mathbf{X}}_{3d}^{\mathrm{CGR}}\\ \hat{\mathbf{Y}}_{3d}^{\mathrm{CGR}}\end{pmatrix}.$$

Notably, the drawback of a separate production of the domain estimates, through composite calibration at the domain level, is the loss of consistency among estimates at population level and domain level.

The above considerations extend to domain estimation for matrix sampling design (d).

6 A simulation study

We have conducted a simulation to study the relative performance of the various composite estimators for the nested version of the basic design (c). Values of correlated scalar variables x and y were generated from a bivariate log-normal distribution with mean and variance parameters (μ_x, μ_y) and (σ_x^2, σ_y^2). With fixed $\mu_x = 3$, $\mu_y = 5$, four combinations of variances (σ_x^2, σ_y^2) (5 and 10) and three values of the correlation $\rho(x, y)$ (0.5, 0.7, 0.9) were considered. Variances $\sigma_x^2 = 5$, $\sigma_x^2 = 10$ imply skewness 2.65 and 4.33, respectively, while variances $\sigma_y^2 = 5$, $\sigma_y^2 = 10$ imply skewness 1.43 and 2.15. For each of these twelve settings, a population of size $N = 1{,}000{,}000$ was created. From each of the twelve populations a simple random sample S of size $n = 5{,}000$ was drawn without replacement, and split into three simple random subsamples (S_1, S_2, S_3) with two different allocations, namely, $(n_1 = 2{,}000, n_2 = 2{,}000, n_3 = 1{,}000)$ and $(n_1 = 1{,}500, n_2 = 1{,}500, n_3 = 2{,}000)$, the second allocation giving larger combined samples $S_1 \cup S_3$ and $S_2 \cup S_3$. Thus, a total of 24 simulation settings were created. For each such setting, we computed the HT estimators of the totals t_x and t_y using the full sample S, as well as the HT estimator of t_x using S_1 and S_3 and the HT estimator of t_y using S_2 and S_3. For the HT estimators based on two subsamples, we employed the simple method for combining two subsamples (Gonzales and Eltinge 2008) by a weighting adjustment involving the probability of selection of a population unit in S_1 or in S_3 and in S_2 or in S_3. In addition, for both t_x and t_y we computed the CGR and COR estimators. Each simulation sampling setting was repeated 10,000 times.

The simulated bias (in percent) of all estimators was smaller than 0.05%, with the exception of two settings involving $\sigma_x^2 = 10$, with associated population skewness of 4.33, where the largest observed values 0.14% and 0.17% correspond to CGR and COR for t_x, respectively, in the sample allocation (2,000, 2,000, 1,000), dropping to 0.10% and 0.13% in the more favorable allocation (1,500, 1,500, 2,000). Thus the relative efficiencies of the estimators are evaluated using their simulated design variances.

Table 6.1 shows the efficiency of the composite estimators CGR and COR relative to the HT estimators that use $S_1 \cup S_3$ and $S_2 \cup S_3$. The measure of this relative efficiency is the percent relative difference of variances [V(CGR)-V(HT)]/V(HT) and [V(COR)-V(HT)]/V(HT). A negative value of this measure indicates the efficiency gain achieved by the two composite estimators. Not shown in Table 6.1, the simulated loss of efficiency of the HT estimators of both t_x and t_y due to not using the full sample S is very close to the nominal loss for SRS, that is, 66.8% for the allocation (2,000, 2,000, 1,000), and 43.1% for the allocation (1,500, 1,500, 2,000).

Table 6.1
Relative differences (in percent) of variances of CGR and COR to HT for x and y, based on 10,000 simulated samples with two different sample allocations.

(n1, n2, n3)	(2,000; 2,000; 1,000)				(1,500; 1,500; 2,000)			
	x		*y*		*x*		*y*	
	CGR	**COR**	**CGR**	**COR**	**CGR**	**COR**	**CGR**	**COR**
$\sigma_x^2 = 5$ $\sigma_y^2 = 5$								
$\rho = 0.5$	-2.24	-6.86	26.39	-6.23	-5.19	-6.29	12.59	-6.52
$\rho = 0.7$	-11.90	-14.75	10.21	-13.96	-12.78	-13.24	0.25	-13.13
$\rho = 0.9$	-24.89	-28.57	-12.49	-28.10	-21.55	-23.37	-14.55	-23.03
$\sigma_x^2 = 5$ $\sigma_y^2 = 10$								
$\rho = 0.5$	-0.27	-6.75	6.50	-6.26	-3.94	-6.60	0.50	-6.44
$\rho = 0.7$	-11.47	-14.56	-6.29	-14.04	-12.87	-13.51	-9.51	-13.10
$\rho = 0.9$	-28.14	-28.42	-25.74	-28.23	-23.70	-23.54	-22.07	-23.09
$\sigma_x^2 = 10$ $\sigma_y^2 = 5$								
$\rho = 0.5$	-4.57	-6.51	28.64	-6.17	-5.90	-5.98	17.57	-6.44
$\rho = 0.7$	-11.29	-14.37	16.08	-13.92	-11.66	-12.90	6.69	-13.00
$\rho = 0.9$	-20.32	-28.09	-2.46	-28.19	-18.46	-22.97	-6.97	-22.91
$\sigma_x^2 = 10$ $\sigma_y^2 = 10$								
$\rho = 0.5$	-4.79	-6.49	8.54	-6.13	-6.06	-6.22	3.41	-6.34
$\rho = 0.7$	-13.27	-14.28	-2.57	-13.95	-13.27	-13.15	-6.00	-12.93
$\rho = 0.9$	-26.01	-28.06	-20.37	-28.21	-22.18	-23.17	-18.48	-22.89

For the variable x, using the CGR estimator at low correlation $\rho = 0.5$ and with allocation (2,000, 2,000, 1,000) leads to an efficiency gain that ranges from 0.27% to 4.79% at the four different variance

settings; this gain reflects the amount of lost information recovered by the CGR estimator. Substantial gain is achieved at $\rho = 0.7$, ranging from 11.29% to 13.27%, and more so at $\rho = 0.9$, ranging from 20.32% to 28.14%. With sample allocation (1,500, 1,500, 2,000) the CGR estimator performs better at $\rho = 0.5$, and $\rho = 0.7$, but not at $\rho = 0.9$. Additional gain is achieved by the COR estimator, which is more efficient than the CGR estimator in all but two settings (where the estimators are equally efficient, see column 7). The efficiency of the COR estimator relative to HT estimator is close to the nominal for SRS efficiency, which is 6.25, 13.92 and 28.12 at $\rho = 0.5$, $\rho = 0.7$, $\rho = 0.9$, respectively, for allocation (2,000, 2,000, 1,000), and 6.417, 13.186 and 23.30 for allocation (1,500, 1,500, 2,000); see quantity E in Section 2, third last paragraph. As expected, the CGR estimator competes better with the COR estimator with increasing correlation and sample size.

For the variable y, the CGR estimator is inferior to the HT estimator at correlation level $\rho = 0.5$ and in half of the simulated settings at $\rho = 0.7$; see positive values in columns 4 and 8. This inefficiency of the CGR estimator ranges from 6.50% (at $\rho = 0.7$) to 28.64% (at $\rho = 0.5$) in the sample allocation (2,000, 2,000, 1,000), and reduces to 0.25% (at $\rho = 0.7$) to 17.57% (at $\rho = 0.5$) in the sample allocation (1,500, 1,500, 2,000). This is explained by the larger skewness of x (the x variable being used a auxiliary to y in the regression procedure); the lower levels of inefficiency are observed at $\sigma_y^2 = 10$, when the differential in skewness between x and y is the smallest. On the other hand, at correlation $\rho = 0.9$ and with allocation (2,000, 2,000, 1,000), the efficiency gain of the CGR estimator relative to the HT estimator ranges from 2.46% (when the skewness differential is the largest) to 25.74% (when the skewness differential is the smallest), with similar efficiency levels displayed for allocation (1,500, 1,500, 2,000). The COR estimator is more efficient than the CGR estimator in all settings, the relative efficiency being close to the nominal one for SRS (same efficiency as with x). For y too, the CGR estimator competes better with COR estimator with increasing correlation and sample size.

This limited empirical study, which essentially simulates the SRS version of Theorem 1(a'), confirms the theory on the efficiency of the optimal estimator COR, even for modest sample size, and shows the usefulness of the two composite estimators CGR and COR in partially recovering the information loss due to splitting the full questionnaire. It also shows that the practical CGR estimator is not always a good substitute of the COR estimator for small samples and low correlation between x and y.

7 Discussion

The proposed estimation method for matrix sampling involves a single-step calibration of the weights of the combined sample. Estimates of totals for all variables can be obtained by using only the units of sample S_3 and their calibrated weights which incorporate all the available information from all three samples. These weights could be used to calculate other weighted statistics, including means, ratios, quantiles and regression coefficients. When the second-order inclusion probabilities are known, including cross-sample inclusion probabilities in the nested case, the calibration procedure of Section 2 can produce composite optimal regression estimators and their variances, but with great computational difficulty. For general sampling settings, the much simpler calibration scheme of Section 3 generates readily composite generalized regression estimators, which for certain sampling strategies are optimal regression estimators.

Estimation of the variance of a CGR estimator may, in principle, be based on the method of Taylor linearization of the generalized regression estimator (see, e.g., Särndal et al. 1992, pages 235, 237). This

approach requires calculations that may not be practical, or even feasible for complex sampling designs because the second-order inclusion probabilities are rarely known. Replication methods for variance estimation, such as the jackknife method or the bootstrap method (see, for example, Rust and Rao 1996), can be applied to the CGR estimators of the previous sections. For example, the jackknife method, customarily used in surveys with stratified multistage sampling design, could be used to replicate the calibration procedures that give rise to the CGR estimators. For the non-nested design, this requires applying the jackknife method to the combined sample, with the three independent samples treated as sample superstrata containing the sample strata. The replication procedure would involve then the combined sample sorted by sample and by strata within each sample, to produce replicates of the calibrated weights defined in the previous sections. The total number of strata used in the jackknife replication procedure is the total number of strata in the three samples, with each replicate involving all strata. Public-use microfiles may include the replicate calibrated weights for easy variance estimation by users. For this purpose too, replicate weights for S_3 only need to be included, bringing about substantial economy of data storage in such microfiles. The case of nested design is more complicated. Further investigation in this direction will be a topic of separate study.

The described estimation method may be readily adapted to matrix sampling designs with more than two subquestionnaires or more than three subsamples, making more evident the operational power of the calibration procedure. In each case, the crucial step is to determine the design matrix $\mathcal{X}$. In such designs there may be more complex patterns with respect to the number of subquestionnaires administered to the various subsamples. All composite estimates can then be obtained using the weighted variable values only from the minimum number of subsamples that in combination contain all items.

Acknowledgements

The author thanks the Editor, Associate Editor and two referees for their comments and suggestions, which have substantially improved the paper.

Appendix

Proof of Lemma 1

For the partitioned matrix $\mathcal{X} = (\mathbf{X}, \boldsymbol{\Psi})$, the vector $\mathbf{c} = \mathbf{w} + \mathbf{R}\mathcal{X}(\mathcal{X}'\mathbf{R}\mathcal{X})^{-1}(\mathbf{t}_{\mathcal{X}} - \mathcal{X}'\mathbf{w})$ takes the form

$$\begin{aligned}\mathbf{c} &= \mathbf{w} + (\mathbf{R}\mathbf{X}, \mathbf{R}\boldsymbol{\Psi})\begin{pmatrix}\mathbf{X}'\mathbf{R}\mathbf{X} & \mathbf{X}'\mathbf{R}\boldsymbol{\Psi}\\ \boldsymbol{\Psi}'\mathbf{R}\mathbf{X} & \boldsymbol{\Psi}'\mathbf{R}\boldsymbol{\Psi}\end{pmatrix}^{-1}\begin{pmatrix}\mathbf{t}_{\mathbf{X}} - \mathbf{X}'\mathbf{w}\\ \mathbf{t}_{\boldsymbol{\Psi}} - \boldsymbol{\Psi}'\mathbf{w}\end{pmatrix}\\ &= \mathbf{w} + (\mathbf{R}\mathbf{X}\mathbf{A}_{11} + \mathbf{R}\boldsymbol{\Psi}\mathbf{A}_{21})(\mathbf{t}_{\mathbf{X}} - \mathbf{X}'\mathbf{w}) + (\mathbf{R}\mathbf{X}\mathbf{A}_{12} + \mathbf{R}\boldsymbol{\Psi}\mathbf{A}_{22})(\mathbf{t}_{\boldsymbol{\Psi}} - \boldsymbol{\Psi}'\mathbf{w}),\end{aligned}$$

where, from algebra of partitioned matrices, $\mathbf{A}_{11} = [\mathbf{X}'\mathbf{R}\mathbf{X} - \mathbf{X}'\mathbf{R}\boldsymbol{\Psi}(\boldsymbol{\Psi}'\mathbf{R}\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}'\mathbf{R}\mathbf{X}]^{-1} = [\mathbf{X}'\mathbf{R}(\mathbf{I} - \mathbf{P}_{\boldsymbol{\Psi}})\mathbf{X}]^{-1}$ with $\mathbf{P}_{\boldsymbol{\Psi}} = \boldsymbol{\Psi}(\boldsymbol{\Psi}'\mathbf{R}\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}'\mathbf{R}$, $\mathbf{A}_{22} = [\boldsymbol{\Psi}'\mathbf{R}(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\boldsymbol{\Psi}']^{-1}$ with $\mathbf{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}'\mathbf{R}\mathbf{X})^{-1}\mathbf{X}'\mathbf{R}$, $\mathbf{A}_{12} = -(\mathbf{X}'\mathbf{R}\mathbf{X})^{-1}$

$(\mathbf{X}'\mathbf{R}\boldsymbol{\Psi})\mathbf{A}_{22}$ and $\mathbf{A}_{21} = -(\boldsymbol{\Psi}'\mathbf{R}\boldsymbol{\Psi})^{-1}(\boldsymbol{\Psi}'\mathbf{R}\mathbf{X})\mathbf{A}_{11}$. Then, equation (2.9) follows without difficulty. To prove equation (2.10), we set $\mathbf{c}_{\boldsymbol{\Psi}} = \mathbf{w} + \mathbf{R}\boldsymbol{\Psi}(\boldsymbol{\Psi}'\mathbf{R}\boldsymbol{\Psi})^{-1}(\mathbf{t}_{\boldsymbol{\Psi}} - \boldsymbol{\Psi}'\mathbf{w})$, so that $(\mathbf{X}'\mathbf{R}\boldsymbol{\Psi})(\boldsymbol{\Psi}'\mathbf{R}\boldsymbol{\Psi})^{-1}(\mathbf{t}_{\boldsymbol{\Psi}} - \boldsymbol{\Psi}'\mathbf{w}) = \mathbf{X}'\mathbf{c}_{\boldsymbol{\Psi}} - \mathbf{X}'\mathbf{w}$, and use the alternative form $\mathbf{A}_{22} = (\boldsymbol{\Psi}'\mathbf{R}\boldsymbol{\Psi})^{-1} + (\boldsymbol{\Psi}'\mathbf{R}\boldsymbol{\Psi})^{-1}(\boldsymbol{\Psi}'\mathbf{R}\mathbf{X})\mathbf{A}_{11}(\mathbf{X}'\mathbf{R}\boldsymbol{\Psi})(\boldsymbol{\Psi}'\mathbf{R}\boldsymbol{\Psi})^{-1}$ to write $\mathbf{c}$ above without the second term as

$$
\begin{aligned}
\mathbf{w} \;+\; & \mathbf{R}\boldsymbol{\Psi}\mathbf{A}_{22}(\mathbf{t}_{\boldsymbol{\Psi}} - \boldsymbol{\Psi}'\mathbf{w}) - \mathbf{R}\mathbf{X}(\mathbf{X}'\mathbf{R}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{R}\boldsymbol{\Psi})\mathbf{A}_{22}(\mathbf{t}_{\boldsymbol{\Psi}} - \boldsymbol{\Psi}'\mathbf{w}) \\
= \; & \mathbf{w} + \left[\mathbf{R}\boldsymbol{\Psi}(\boldsymbol{\Psi}'\mathbf{R}\boldsymbol{\Psi})^{-1} + \mathbf{R}\boldsymbol{\Psi}(\boldsymbol{\Psi}'\mathbf{R}\boldsymbol{\Psi})^{-1}(\boldsymbol{\Psi}'\mathbf{R}\mathbf{X})\mathbf{A}_{11}(\mathbf{X}'\mathbf{R}\boldsymbol{\Psi})(\boldsymbol{\Psi}'\mathbf{R}\boldsymbol{\Psi})^{-1}\right](\mathbf{t}_{\boldsymbol{\Psi}} - \boldsymbol{\Psi}'\mathbf{w}) \\
- \; & \mathbf{R}\mathbf{X}(\mathbf{X}'\mathbf{R}\mathbf{X})^{-1}\left[\mathbf{I} + (\mathbf{X}'\mathbf{R}\boldsymbol{\Psi})(\boldsymbol{\Psi}'\mathbf{R}\boldsymbol{\Psi})^{-1}(\boldsymbol{\Psi}'\mathbf{R}\mathbf{X})\mathbf{A}_{11}\right](\mathbf{X}'\mathbf{R}\boldsymbol{\Psi})(\boldsymbol{\Psi}'\mathbf{R}\boldsymbol{\Psi})^{-1}(\mathbf{t}_{\boldsymbol{\Psi}} - \boldsymbol{\Psi}'\mathbf{w}) \\
= \; & \mathbf{c}_{\boldsymbol{\Psi}} + \mathbf{R}\boldsymbol{\Psi}(\boldsymbol{\Psi}'\mathbf{R}\boldsymbol{\Psi})^{-1}(\boldsymbol{\Psi}'\mathbf{R}\mathbf{X})\mathbf{A}_{11}(\mathbf{X}'\mathbf{c}_{\boldsymbol{\Psi}} - \mathbf{X}'\mathbf{w}) \\
- \; & \mathbf{R}\mathbf{X}(\mathbf{X}'\mathbf{R}\mathbf{X})^{-1}\left[\mathbf{I} + (\mathbf{X}'\mathbf{R}\boldsymbol{\Psi})(\boldsymbol{\Psi}'\mathbf{R}\boldsymbol{\Psi})^{-1}(\boldsymbol{\Psi}'\mathbf{R}\mathbf{X})\mathbf{A}_{11}\right](\mathbf{X}'\mathbf{c}_{\boldsymbol{\Psi}} - \mathbf{X}'\mathbf{w}) \\
= \; & \mathbf{c}_{\boldsymbol{\Psi}} + \mathbf{R}\boldsymbol{\Psi}(\boldsymbol{\Psi}'\mathbf{R}\boldsymbol{\Psi})^{-1}(\boldsymbol{\Psi}'\mathbf{R}\mathbf{X})\mathbf{A}_{11}(\mathbf{X}'\mathbf{c}_{\boldsymbol{\Psi}} - \mathbf{X}'\mathbf{w}) \\
- \; & \mathbf{R}\mathbf{X}(\mathbf{X}'\mathbf{R}\mathbf{X})^{-1}\left[\mathbf{I} + (\mathbf{X}'\mathbf{R}\mathbf{X} - \mathbf{A}_{11}^{-1})\mathbf{A}_{11}\right](\mathbf{X}'\mathbf{c}_{\boldsymbol{\Psi}} - \mathbf{X}'\mathbf{w}) \\
= \; & \mathbf{c}_{\boldsymbol{\Psi}} + \left[\mathbf{R}\boldsymbol{\Psi}(\boldsymbol{\Psi}'\mathbf{R}\boldsymbol{\Psi})^{-1}(\boldsymbol{\Psi}'\mathbf{R}\mathbf{X}) - \mathbf{R}\mathbf{X}\right]\mathbf{A}_{11}(\mathbf{X}'\mathbf{c}_{\boldsymbol{\Psi}} - \mathbf{X}'\mathbf{w}) \\
= \; & \mathbf{c}_{\boldsymbol{\Psi}} - \mathbf{R}(\mathbf{I} - \mathbf{P}_{\boldsymbol{\Psi}})\mathbf{X}\left[\mathbf{X}'\mathbf{R}(\mathbf{I} - \mathbf{P}_{\boldsymbol{\Psi}})\mathbf{X}\right]^{-1}(\mathbf{X}'\mathbf{c}_{\boldsymbol{\Psi}} - \mathbf{X}'\mathbf{w}).
\end{aligned}
$$

Adding to this the second term of $\mathbf{c}$ from (2.9) gives (2.10), in the explicit form

$$\mathbf{c}_{\boldsymbol{\Psi}} + \mathbf{R}(\mathbf{I} - \mathbf{P}_{\boldsymbol{\Psi}})\mathbf{X}\left[\mathbf{X}'\mathbf{R}(\mathbf{I} - \mathbf{P}_{\boldsymbol{\Psi}})\mathbf{X}\right]^{-1}(\mathbf{t}_{\mathbf{X}} - \mathbf{X}'\mathbf{c}_{\boldsymbol{\Psi}}).$$

Proof of Theorem 1

(a) Calibration with design matrix $\mathcal{Z} = (\mathcal{X}, \mathbf{D})$ and vector of totals $\mathbf{t}_{\mathcal{Z}} = (\mathbf{0}', \mathbf{N}')'$, with $\mathbf{0} = (\mathbf{0}', \mathbf{0}')'$, $\mathbf{N} = (\mathbf{N}_1', \mathbf{N}_2', \mathbf{N}_3')'$, gives the vector of calibrated weights $\mathbf{c} = \mathbf{w} + \boldsymbol{\Lambda}\mathcal{Z}(\mathcal{Z}'\boldsymbol{\Lambda}\mathcal{Z})^{-1}(\mathbf{t}_{\mathcal{Z}} - \mathcal{Z}'\mathbf{w})$, which by Lemma 1 is written as $\mathbf{c} = \mathbf{c}_{\mathbf{D}} + \mathbf{L}_{\mathbf{D}}\mathcal{X}(\mathcal{X}'\mathbf{L}_{\mathbf{D}}\mathcal{X})^{-1}(\mathbf{0} - \mathcal{X}'\mathbf{c}_{\mathbf{D}})$, where $\mathbf{c}_{\mathbf{D}} = \mathbf{w} + \boldsymbol{\Lambda}\mathbf{D}(\mathbf{D}'\boldsymbol{\Lambda}\mathbf{D})^{-1}(\mathbf{N} - \mathbf{D}'\mathbf{w})$ and $\mathbf{L}_{\mathbf{D}} = \boldsymbol{\Lambda}(\mathbf{I} - \mathbf{P}_{\mathbf{D}})$, with $\mathbf{P}_{\mathbf{D}} = \mathbf{D}(\mathbf{D}'\boldsymbol{\Lambda}\mathbf{D})^{-1}\mathbf{D}'\boldsymbol{\Lambda}$. For STRSRS with $f_{ih} = n_{ih}/N_{ih}$, $\mathbf{D}'\mathbf{w} = \hat{\mathbf{N}} = \mathbf{N}$, and thus $\mathbf{c} = \mathbf{w} + \mathbf{L}_{\mathbf{D}}\mathcal{X}(\mathcal{X}'\mathbf{L}_{\mathbf{D}}\mathcal{X})^{-1}(\mathbf{0} - \mathcal{X}'\mathbf{w})$. Then, in view of (2.8), in order to show that $\hat{\boldsymbol{\mathcal{B}}} = \hat{\boldsymbol{\mathcal{B}}}^{o}$ it suffices to show that $\mathbf{L}_{\mathbf{D}} = \boldsymbol{\Lambda}^{0}$. For STRSRS it is easy to show that $\boldsymbol{\Lambda}^{0} = \mathrm{diag}\{\lambda_{ih}(\mathbf{I} - \mathbf{P}_{\mathbf{1}ih})\}$, where $\lambda_{ih} = N_{ih}^{2}(1 - f_{ih})/[n_{ih}(n_{ih} - 1)]$ and $\mathbf{P}_{\mathbf{1}ih} = \mathbf{1}_{ih}(\mathbf{1}_{ih}'\mathbf{1}_{ih})^{-1}\mathbf{1}_{ih}'$. Next, observe that the matrix $\mathbf{P}_{\mathbf{D}}$ is diagonal with ih^{th} entry $\mathbf{1}_{ih}(\mathbf{1}_{ih}'\boldsymbol{\Lambda}_{ih}\mathbf{1}_{ih})^{-1}\mathbf{1}_{ih}'\boldsymbol{\Lambda}_{ih} = \mathbf{P}_{\mathbf{1}ih}$, because the elements of $\boldsymbol{\Lambda}_{ih}$ are constant. Since this constant element is $w_{ik}/q_{ik} = (N_{ih}/n_{ih})[N_{ih}(1 - f_{ih})/(n_{ih} - 1)] = \lambda_{ih}$, we get $\mathbf{L}_{\mathbf{D}} = \mathrm{diag}\{\boldsymbol{\Lambda}_{ih}(\mathbf{I} - \mathbf{P}_{\mathbf{1}ih})\} = \boldsymbol{\Lambda}^{0}$, o.e.d.

(b) For Poisson sampling, $\boldsymbol{\Lambda}_{i}^{0} = \mathrm{diag}\{(1 - \pi_{ihk})/\pi_{ihk}^{2}\}, h = 1, \ldots, H_{i}$. The proof follows immediately upon observing that with the specified constants q_{ik} in the entries of $\boldsymbol{\Lambda}_{i}$ we have $\boldsymbol{\Lambda}_{i} = \boldsymbol{\Lambda}_{i}^{0}$.

(a') For simplicity drop the stratum subscript. Simple random subsampling is done sequentially with fixed sizes n_1, n_2 and n_3. It can be shown that the first-and-second order marginal inclusion

probabilities for S_i are $\pi_{ik} = n_i/N$ and $\pi_{ikl} = n_i(n_i-1)/[N(N-1)]$, as if S_i was drawn directly from U. A combinatorial argument shows that the conditional (given S) second-order inclusion probability for S_i and S_j is $\pi_{ikjl|S} = n_i n_j/[n(n-1)]$ and thus the marginal inclusion probability is $\pi_{ikjl} = n_i n_j/[N(N-1)]$. For $k = l, \pi_{ikjk} = 0$. Then $\Delta_{kl} = \pi_{ikjl} - \pi_{ik}\pi_{jl} = n_i n_j/[N^2(N-1)]$ and $\Delta_{kk} = -n_i n_j/N^2$. Thus $\Delta_{kl} \approx 0$, for $k, l \in U$ when the sampling fractions are small, and then $\boldsymbol{\Lambda}^0 \approx \text{diag}\{\boldsymbol{\Lambda}_i^0\}$. Optimality of the CGR then follows from Theorem 1 (a).

(*b*') Randomly assigning the units of S to three subsamples, with fixed expected subsample size, implies that inclusion of the units is done independently within and between the subsamples. Since in Poisson sampling the units of U are also included in S independently, $\Delta_{kl} = \pi_{ikjl} - \pi_{ik}\pi_{jl} = 0$ and $\Delta_{kk} = -\pi_{ik}\pi_{jl}$. Δ_{kk} is approximately zero for small sampling fractions, and then $\boldsymbol{\Lambda}^0 \approx \text{diag}\{\boldsymbol{\Lambda}_i^0\}$. Optimality of the CGR follows then from Theorem 1(*b*).

Proof of Theorem 2

We start with the expression of the CGR estimator. By Lemma 1, with partitioned design matrix $(\boldsymbol{\mathcal{X}}, \mathbf{Z})$ and $\mathbf{R} = \boldsymbol{\Lambda}$, the calibrated weight vector $\mathbf{c}$ can be written as $\mathbf{c} = \mathbf{c}_\mathbf{Z} + \mathbf{L}_\mathbf{Z}\boldsymbol{\mathcal{X}}(\boldsymbol{\mathcal{X}}'\mathbf{L}_\mathbf{Z}\boldsymbol{\mathcal{X}})^{-1}(\mathbf{0} - \boldsymbol{\mathcal{X}}'\mathbf{c}_\mathbf{Z})$, where $\mathbf{c}_\mathbf{Z} = \mathbf{w} + \boldsymbol{\Lambda}\mathbf{Z}(\mathbf{Z}'\boldsymbol{\Lambda}\mathbf{Z})^{-1}(\mathbf{t}_{(\mathbf{z})} - \mathbf{Z}'\mathbf{w})$ and $\mathbf{L}_\mathbf{Z} = \boldsymbol{\Lambda}(\mathbf{I} - \mathbf{P}_\mathbf{Z})$. Then $\hat{\boldsymbol{\mathcal{X}}}_3^{\text{GR}} = \boldsymbol{\mathcal{X}}_3'\mathbf{c}_\mathbf{Z} = \hat{\boldsymbol{\mathcal{X}}}_3 + \boldsymbol{\mathcal{X}}_3'\boldsymbol{\Lambda}\mathbf{Z}(\mathbf{Z}'\boldsymbol{\Lambda}\mathbf{Z})^{-1}(\mathbf{t}_{(\mathbf{z})} - \hat{\mathbf{Z}})$ and $\hat{\boldsymbol{\mathcal{X}}}^{\text{GR}} = \hat{\boldsymbol{\mathcal{X}}} + \boldsymbol{\mathcal{X}}'\boldsymbol{\Lambda}\mathbf{Z}(\mathbf{Z}'\boldsymbol{\Lambda}\mathbf{Z})^{-1}(\mathbf{t}_{(\mathbf{z})} - \hat{\mathbf{Z}})$. It follows that the CGR estimator is given by $\boldsymbol{\mathcal{X}}_3'\mathbf{c} = \hat{\boldsymbol{\mathcal{X}}}_3^{\text{GR}} - \hat{\boldsymbol{\mathcal{B}}}\hat{\boldsymbol{\mathcal{X}}}^{\text{GR}}$, where $\hat{\boldsymbol{\mathcal{B}}} = [\boldsymbol{\mathcal{X}}_3'\boldsymbol{\Lambda}(\mathbf{I} - \mathbf{P}_\mathbf{Z})\boldsymbol{\mathcal{X}}][\boldsymbol{\mathcal{X}}'\boldsymbol{\Lambda}(\mathbf{I} - \mathbf{P}_\mathbf{Z})\boldsymbol{\mathcal{X}}]^{-1}$.

(a) Since $\mathbf{P}_\mathbf{Z} = \text{diag}\{\mathbf{P}_{\mathbf{Z}_i}\}$ and, for SRS, $\boldsymbol{\Lambda}^0 = \text{diag}\{\lambda_i(\mathbf{I} - \mathbf{P}_{\mathbf{1}i})\}$, where $\lambda_i = N^2(1 - f_i)/[n_i(n_i - 1)]$ and $\mathbf{P}_{\mathbf{1}i} = \mathbf{1}_i(\mathbf{1}_i'\mathbf{1}_i)^{-1}\mathbf{1}_i'$, we have $\boldsymbol{\Lambda}^0(\mathbf{I} - \mathbf{P}_\mathbf{Z}) = \text{diag}\{\lambda_i(\mathbf{I} - \mathbf{P}_{\mathbf{1}i})(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_i})\}$. Now, by assumption $\mathbf{1} = \mathbf{Z}_i\mathbf{h}_i$, so that $\mathbf{1}'\mathbf{P}_{\mathbf{Z}_i} = \mathbf{1}'$ and hence $\mathbf{P}_{\mathbf{1}i}(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_i}) = \mathbf{0}$. It follows that $\boldsymbol{\Lambda}^0(\mathbf{I} - \mathbf{P}_\mathbf{Z}) = \text{diag}\{\lambda_i(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_i})\}$ and, since the matrices $\mathbf{I} - \mathbf{P}_{\mathbf{Z}_i}$ are idempotent, $(\mathbf{I} - \mathbf{P}_\mathbf{Z})'\boldsymbol{\Lambda}^0(\mathbf{I} - \mathbf{P}_\mathbf{Z}) = \text{diag}\{\lambda_i(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_i})\}$. But $\lambda_i = w_{ik}/q_{ik}$, where $w_{ik} = N/n_i$ and q_{ik} are the specified constants in the entries of $\boldsymbol{\Lambda}_i$. It follows that $(\mathbf{I} - \mathbf{P}_\mathbf{Z})'\boldsymbol{\Lambda}^0(\mathbf{I} - \mathbf{P}_\mathbf{Z}) = \text{diag}\{\boldsymbol{\Lambda}_i(\mathbf{I} - \mathbf{P}_{\mathbf{Z}_i})\} = \boldsymbol{\Lambda}(\mathbf{I} - \mathbf{P}_\mathbf{Z})$ and thus $\hat{\boldsymbol{\mathcal{B}}} = \hat{\boldsymbol{\mathcal{B}}}^{wo}$, so that $\hat{\boldsymbol{\mathcal{X}}}_3^{\text{GR}} - \hat{\boldsymbol{\mathcal{B}}}\hat{\boldsymbol{\mathcal{X}}}^{\text{GR}} = \hat{\boldsymbol{\mathcal{X}}}_3^{\text{GR}} - \hat{\boldsymbol{\mathcal{B}}}^{wo}\hat{\boldsymbol{\mathcal{X}}}^{\text{GR}}$.

(b) By Lemma 1, with the partitioned design matrix $\boldsymbol{\mathcal{Z}} = (\boldsymbol{\mathcal{X}}, \mathbf{Z}, \mathbf{D})$ and vector of totals $\mathbf{t}_\mathcal{Z} = (\mathbf{0}', \mathbf{t}_{(\mathbf{z})}', \mathbf{N}')'$, the vector of calibrated weights $\mathbf{c} = \mathbf{w} + \boldsymbol{\Lambda}\boldsymbol{\mathcal{Z}}(\boldsymbol{\mathcal{Z}}'\boldsymbol{\Lambda}\boldsymbol{\mathcal{Z}})^{-1}(\mathbf{t}_\mathcal{Z} - \boldsymbol{\mathcal{Z}}'\mathbf{w})$ can be written as $\mathbf{c} = \mathbf{c}_\mathbf{D} + \mathbf{L}_\mathbf{D}(\boldsymbol{\mathcal{X}}, \mathbf{Z})\left[(\boldsymbol{\mathcal{X}}, \mathbf{Z})'\mathbf{L}_\mathbf{D}(\boldsymbol{\mathcal{X}}, \mathbf{Z})\right]^{-1}\left[(\mathbf{0}', \mathbf{t}_{(\mathbf{z})}')' - (\boldsymbol{\mathcal{X}}, \mathbf{Z})'\mathbf{c}_\mathbf{D}\right]$, where $\mathbf{c}_\mathbf{D} = \mathbf{w} + \boldsymbol{\Lambda}\mathbf{D}(\mathbf{D}'\boldsymbol{\Lambda}\mathbf{D})^{-1}(\mathbf{N} - \mathbf{D}'\mathbf{w})$ and $\mathbf{L}_\mathbf{D} = \boldsymbol{\Lambda}(\mathbf{I} - \mathbf{P}_\mathbf{D})$, with $\mathbf{P}_\mathbf{D} = \mathbf{D}(\mathbf{D}'\boldsymbol{\Lambda}\mathbf{D})^{-1}\mathbf{D}'\boldsymbol{\Lambda}$. But, as shown in the proof of Theorem 1(a), $\mathbf{c}_\mathbf{D} = \mathbf{w}$ and $\mathbf{L}_\mathbf{D} = \boldsymbol{\Lambda}^0$. Thus, $\mathbf{c} = \mathbf{w} + \boldsymbol{\Lambda}^0(\boldsymbol{\mathcal{X}}, \mathbf{Z})\left[(\boldsymbol{\mathcal{X}}, \mathbf{Z})'\boldsymbol{\Lambda}^0(\boldsymbol{\mathcal{X}}, \mathbf{Z})\right]^{-1}\left[(\mathbf{0}', \mathbf{t}_{(\mathbf{z})}')' - (\boldsymbol{\mathcal{X}}, \mathbf{Z})'\mathbf{w}\right]$. Next, by applying again Lemma 1, now with $\mathbf{R} = \boldsymbol{\Lambda}^0$ and design matrix $(\boldsymbol{\mathcal{X}}, \mathbf{Z})$, we get $\mathbf{c} = \mathbf{c}_\mathbf{Z} + \mathbf{L}_\mathbf{Z}^0\boldsymbol{\mathcal{X}}(\boldsymbol{\mathcal{X}}'\mathbf{L}_\mathbf{Z}^0\boldsymbol{\mathcal{X}})^{-1}(\mathbf{0} - \boldsymbol{\mathcal{X}}'\mathbf{c}_\mathbf{Z})$, where $\mathbf{c}_\mathbf{Z} = \mathbf{w} + \boldsymbol{\Lambda}^0\mathbf{Z}(\mathbf{Z}'\boldsymbol{\Lambda}^0\mathbf{Z})^{-1}$

$(\mathbf{t}_{(\mathbf{z})} - \mathbf{Z}'\mathbf{w})$ and $\mathbf{L}_{\mathbf{Z}}^0 = \mathbf{\Lambda}^0(\mathbf{I} - \mathbf{P}_{\mathbf{Z}}^0)$. Then it follows that the CGR estimator is $\mathcal{X}_3'\mathbf{c} = \mathcal{X}_3'\mathbf{c}_{\mathbf{Z}} - \mathcal{X}_3'\mathbf{L}_{\mathbf{Z}}^0\mathcal{X}(\mathcal{X}'\mathbf{L}_{\mathbf{Z}}^0\mathcal{X})^{-1}\mathcal{X}'\mathbf{c}_{\mathbf{Z}} = \hat{\mathcal{X}}_3^{\mathrm{OR}} - \hat{\mathcal{B}}^o\hat{\mathcal{X}}^{\mathrm{OR}}$, in obvious expressions for $\hat{\mathcal{X}}_3^{\mathrm{OR}}, \hat{\mathcal{X}}^{\mathrm{OR}}$ and $\hat{\mathcal{B}}^o$.

(c) It was shown in the proof of Theorem 1 that $\mathbf{\Lambda} = \mathbf{\Lambda}^0$. Clearly then it holds that $\hat{\mathcal{X}}_3^{\mathrm{GR}} = \hat{\mathcal{X}}_3^{\mathrm{OR}}$, $\hat{\mathcal{X}}^{\mathrm{GR}} = \hat{\mathcal{X}}^{\mathrm{OR}}$ and $\hat{\mathcal{B}} = \hat{\mathcal{B}}^o$, and thus $\hat{\mathcal{X}}_3^{\mathrm{GR}} - \hat{\mathcal{B}}\hat{\mathcal{X}}^{\mathrm{GR}} = \hat{\mathcal{X}}_3^{\mathrm{OR}} - \hat{\mathcal{B}}^o\hat{\mathcal{X}}^{\mathrm{OR}}$.

Proof of Proposition 1

All matrices appearing in this proof are defined at the population level. Partitioning the matrix $\mathcal{X}$ in (4.4) as $(\mathbf{Z}, \mathbf{\Psi})$, where $\mathbf{Z}$ consists of the second and fourth columns, and $\mathbf{\Psi}$ of the rest, and applying Lemma 1 with $\mathbf{R} = \mathbf{\Lambda}^0 = \{(\pi_{kl} - \pi_k\pi_l)/\pi_k\pi_l\}$, we obtain the vector of calibrated weights decomposed as

$$\mathbf{c} = \mathbf{w} + \mathbf{L}_{\mathbf{\Psi}}^0\mathbf{Z}(\mathbf{Z}'\mathbf{L}_{\mathbf{\Psi}}^0\mathbf{Z})^{-1}[\mathbf{0} - \mathbf{Z}'\mathbf{w}] + \mathbf{L}_{\mathbf{Z}}^0\mathbf{\Psi}(\mathbf{\Psi}'\mathbf{L}_{\mathbf{Z}}^0\mathbf{\Psi})^{-1}[\mathbf{0} - \mathbf{\Psi}'\mathbf{w}],$$

where $\mathbf{L}_{\mathbf{Z}}^0 = \mathbf{\Lambda}^0(\mathbf{I} - \mathbf{P}_{\mathbf{Z}}^0)$ with $\mathbf{P}_{\mathbf{Z}}^0 = \mathbf{Z}(\mathbf{Z}'\mathbf{\Lambda}^0\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{\Lambda}^0$. The estimator $\hat{\mathbf{Z}}^B$ in (4.2) is obtained as $\mathbf{Z}_{3-}'\mathbf{c}$, where $\mathbf{Z}_{3-} = (\mathbf{0}', \mathbf{0}', \mathbf{Z}_3')'$. The last two terms of (4.2) are consolidated in the term $\mathbf{Z}_{3-}'\mathbf{L}_{\mathbf{Z}}^0\mathbf{\Psi}(\mathbf{\Psi}'\mathbf{L}_{\mathbf{Z}}^0\mathbf{\Psi})^{-1}[\mathbf{0} - \mathbf{\Psi}'\mathbf{w}]$. These two terms vanish only if $\mathbf{Z}_{3-}'\mathbf{L}_{\mathbf{Z}}^0\mathbf{\Psi}(= \mathbf{Z}_{3-}'\mathbf{\Lambda}^0\mathbf{\Psi} - \mathbf{Z}_{3-}'\mathbf{\Lambda}^0\mathbf{Z}(\mathbf{Z}'\mathbf{\Lambda}^0\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{\Lambda}^0\mathbf{\Psi}) = \mathbf{0}$. First, we easily get $\mathbf{Z}_{3-}'\mathbf{\Lambda}^0\mathbf{\Psi} = (\mathbf{Z}_3'\mathbf{\Lambda}_3^0\mathbf{X}_3, \mathbf{Z}_3'\mathbf{\Lambda}_3^0\mathbf{Y}_3)$ and $\mathbf{Z}_{3-}'\mathbf{\Lambda}^0\mathbf{Z} = \mathbf{Z}_3'\mathbf{\Lambda}_3^0\mathbf{Z}_3(\mathbf{I}, \mathbf{I})$, as well as

$$\mathbf{Z}'\mathbf{\Lambda}^0\mathbf{\Psi} = \begin{pmatrix} \mathbf{Z}_1'\mathbf{\Lambda}_1^0\mathbf{X}_1 + \mathbf{Z}_3'\mathbf{\Lambda}_3^0\mathbf{X}_3 & \mathbf{Z}_3'\mathbf{\Lambda}_3^0\mathbf{Y}_3 \\ \mathbf{Z}_3'\mathbf{\Lambda}_3^0\mathbf{X}_3 & \mathbf{Z}_2'\mathbf{\Lambda}_2^0\mathbf{Y}_2 + \mathbf{Z}_3'\mathbf{\Lambda}_3^0\mathbf{Y}_3 \end{pmatrix},$$

and

$$\mathbf{Z}'\mathbf{\Lambda}^0\mathbf{Z} = \begin{pmatrix} \mathbf{Z}_1'\mathbf{\Lambda}_1^0\mathbf{Z}_1 + \mathbf{Z}_3'\mathbf{\Lambda}_3^0\mathbf{Z}_3 & \mathbf{Z}_3'\mathbf{\Lambda}_3^0\mathbf{Z}_3 \\ \mathbf{Z}_3'\mathbf{\Lambda}_3^0\mathbf{Z}_3 & \mathbf{Z}_2'\mathbf{\Lambda}_2^0\mathbf{Z}_2 + \mathbf{Z}_3'\mathbf{\Lambda}_3^0\mathbf{Z}_3 \end{pmatrix}.$$

Next we write

$$(\mathbf{Z}'\mathbf{\Lambda}^0\mathbf{Z})^{-1} = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}' & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{A}^{-1} + \mathbf{F}\mathbf{E}^{-1}\mathbf{F}' & -\mathbf{F}\mathbf{E}^{-1} \\ -\mathbf{E}^{-1}\mathbf{F}' & \mathbf{E}^{-1} \end{pmatrix},$$

where $\mathbf{E} = \mathbf{D} - \mathbf{B}'\mathbf{A}^{-1}\mathbf{B}$ and $\mathbf{F} = \mathbf{A}^{-1}\mathbf{B}$. It follows then that $\mathbf{Z}_{3-}'\mathbf{\Lambda}^0\mathbf{Z}(\mathbf{Z}'\mathbf{\Lambda}^0\mathbf{Z})^{-1} = (\mathbf{B}\mathbf{A}^{-1} + \mathbf{B}\mathbf{F}\mathbf{E}^{-1}\mathbf{F}' - \mathbf{B}\mathbf{E}^{-1}\mathbf{F}', \mathbf{B}(\mathbf{I} - \mathbf{F})\mathbf{E}^{-1}) = ((\mathbf{D} - \mathbf{B})\mathbf{E}^{-1}\mathbf{F}', \mathbf{B}(\mathbf{I} - \mathbf{F})\mathbf{E}^{-1})$. Using the analytic expressions $\mathbf{B} = \mathbf{Z}_3'\mathbf{\Lambda}_3^0\mathbf{Z}_3$, $\mathbf{D} = \mathbf{Z}_2'\mathbf{\Lambda}_2^0\mathbf{Z}_2 + \mathbf{Z}_3'\mathbf{\Lambda}_3^0\mathbf{Z}_3$, $\mathbf{F} = (\mathbf{Z}_1'\mathbf{\Lambda}_1^0\mathbf{Z}_1 + \mathbf{Z}_3'\mathbf{\Lambda}_3^0\mathbf{Z}_3)^{-1}\mathbf{Z}_3'\mathbf{\Lambda}_3^0\mathbf{Z}_3$ and $\mathbf{E} = \mathbf{Z}_2'\mathbf{\Lambda}_2^0\mathbf{Z}_2 + \mathbf{Z}_1'\mathbf{\Lambda}_1^0\mathbf{Z}_1\mathbf{F}$, we obtain after some algebra

$$\mathbf{Z}_{3-}'\mathbf{\Lambda}^0\mathbf{Z}(\mathbf{Z}'\mathbf{\Lambda}^0\mathbf{Z})^{-1} = \mathbf{K}^{-1}\left[(\mathbf{Z}_1'\mathbf{\Lambda}_1^0\mathbf{Z}_1)^{-1}, (\mathbf{Z}_2'\mathbf{\Lambda}_2^0\mathbf{Z}_2)^{-1}\right],$$

where $\mathbf{K} = (\mathbf{Z}_1'\mathbf{\Lambda}_1^0\mathbf{Z}_1)^{-1} + (\mathbf{Z}_2'\mathbf{\Lambda}_2^0\mathbf{Z}_2)^{-1} + (\mathbf{Z}_3'\mathbf{\Lambda}_3^0\mathbf{Z}_3)^{-1}$. We can now obtain without much difficulty

$$\begin{aligned} \mathbf{Z}_{3-}'\mathbf{L}_{\mathbf{Z}}^0\mathbf{\Psi} &= \mathbf{Z}_{3-}'\mathbf{\Lambda}^0\mathbf{\Psi} - \mathbf{Z}_{3-}'\mathbf{\Lambda}^0\mathbf{Z}(\mathbf{Z}'\mathbf{\Lambda}^0\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{\Lambda}^0\mathbf{\Psi} \\ &= \mathbf{K}^{-1}\left[(\mathbf{Z}_3'\mathbf{\Lambda}_3^0\mathbf{Z}_3)^{-1}\mathbf{Z}_3'\mathbf{\Lambda}_3^0\mathbf{X}_3 - (\mathbf{Z}_1'\mathbf{\Lambda}_1^0\mathbf{Z}_1)^{-1}\mathbf{Z}_1'\mathbf{\Lambda}_1^0\mathbf{X}_1,\right. \\ &\quad \left.(\mathbf{Z}_3'\mathbf{\Lambda}_3^0\mathbf{Z}_3)^{-1}\mathbf{Z}_3'\mathbf{\Lambda}_3^0\mathbf{Y}_3 - (\mathbf{Z}_2'\mathbf{\Lambda}_2^0\mathbf{Z}_2)^{-1}\mathbf{Z}_2'\mathbf{\Lambda}_2^0\mathbf{Y}_2\right]. \end{aligned}$$

It follows that $\mathbf{Z}'_{3-}\mathbf{L}^0_{\mathbf{z}}\boldsymbol{\Psi} = (\mathbf{0},\mathbf{0})$ only if $\left(\mathbf{Z}'_3\boldsymbol{\Lambda}^0_3\mathbf{Z}_3\right)^{-1}\mathbf{Z}'_3\boldsymbol{\Lambda}^0_3\mathbf{X}_3 = \left(\mathbf{Z}'_1\boldsymbol{\Lambda}^0_1\mathbf{Z}_1\right)^{-1}\mathbf{Z}'_1\boldsymbol{\Lambda}^0_1\mathbf{X}_1$ and $\left(\mathbf{Z}'_3\boldsymbol{\Lambda}^0_3\mathbf{Z}_3\right)^{-1}$ $\mathbf{Z}'_3\boldsymbol{\Lambda}^0_3\mathbf{Y}_3 = \left(\mathbf{Z}'_2\boldsymbol{\Lambda}^0_2\mathbf{Z}_2\right)^{-1}\mathbf{Z}'_2\boldsymbol{\Lambda}^0_2\mathbf{Y}_2$. But these two equations are identical to the equations in (4.6). Since all the matrices in $\left(\mathbf{Z}'_i\boldsymbol{\Lambda}^0_i\mathbf{Z}_i\right)^{-1}\mathbf{Z}'_i\boldsymbol{\Lambda}^0_i\mathbf{X}_i$ are defined at the population level, with the subscript $i = 1,3$ indicating survey, this quantity is constant across surveys only if the design-specific matrix $\boldsymbol{\Lambda}^0_i$ is constant, or if $\boldsymbol{\Lambda}^0_i$ differs among surveys by a constant multiple (depending on the sample size). This holds true also for $\left(\mathbf{Z}'_i\boldsymbol{\Lambda}^0_i\mathbf{Z}_i\right)^{-1}\mathbf{Z}'_i\boldsymbol{\Lambda}^0_i\mathbf{Y}_i$, $i = 2,3$. This completes the proof.

Proof of Proposition 2

Under the sampling scheme (a) of Theorem 1, composite calibration at population level with design matrix $\mathcal{Z} = (\mathcal{X},\mathbf{D})$ and vector of totals $\mathbf{t}_{\mathcal{Z}} = (\mathbf{0}',\mathbf{N}')'$ produces the joint CGR domain estimator of $(\mathbf{t}'_{\mathbf{x}d},\mathbf{t}'_{\mathbf{y}d})'$ based on the weights of S_3 and written in the form $\hat{\mathcal{X}}^{\mathrm{CGR}}_{3d} = \hat{\mathcal{X}}_{3d} + \hat{\mathcal{B}}_d\left(\mathbf{t}_{\mathcal{Z}} - \hat{\mathcal{Z}}\right)$, where $\hat{\mathcal{B}}_d = \mathcal{X}'_{3d}\boldsymbol{\Lambda}\mathcal{Z}\left(\mathcal{Z}'\boldsymbol{\Lambda}\mathcal{Z}\right)^{-1}$. The associated matrix of regression residuals is $\mathcal{X}_{3d} - \mathcal{Z}\hat{\mathcal{B}}'_d$, alternatively written as $(\mathbf{I} - \mathbf{P}_{\mathcal{Z}})\mathcal{X}_{3d}$, with $\mathbf{P}_{\mathcal{Z}} = \mathcal{Z}\left(\mathcal{Z}'\boldsymbol{\Lambda}\mathcal{Z}\right)^{-1}\mathcal{Z}'\boldsymbol{\Lambda}$. Then $\widehat{\mathrm{AV}}\left(\hat{\mathcal{X}}^{\mathrm{CGR}}_{3d}\right) = \mathcal{X}'_{3d}\left(\mathbf{I} - \mathbf{P}_{\mathcal{Z}}\right)'$ $\boldsymbol{\Lambda}^0\left(\mathbf{I} - \mathbf{P}_{\mathcal{Z}}\right)\mathcal{X}_{3d}$. Next recall from the proof of Theorem 1 that $\boldsymbol{\Lambda}^0 = \boldsymbol{\Lambda}(\mathbf{I} - \mathbf{P}_{\mathbf{D}})$, with $\mathbf{P}_{\mathbf{D}} = \mathbf{D}\left(\mathbf{D}'\boldsymbol{\Lambda}\mathbf{D}\right)^{-1}\mathbf{D}'\boldsymbol{\Lambda}$, and notice that $\mathbf{D} = \mathcal{Z}\mathbf{H}$ for a suitable constant matrix $\mathbf{H}$. It is easy to verify that $\mathbf{P}_{\mathbf{D}}\mathbf{P}_{\mathcal{Z}} = \mathbf{P}_{\mathbf{D}}$. It follows then that $\boldsymbol{\Lambda}^0\left(\mathbf{I} - \mathbf{P}_{\mathcal{Z}}\right) = \boldsymbol{\Lambda}\left(\mathbf{I} - \mathbf{P}_{\mathcal{Z}}\right)$ and $\left(\mathbf{I} - \mathbf{P}_{\mathcal{Z}}\right)'\boldsymbol{\Lambda}^0\left(\mathbf{I} - \mathbf{P}_{\mathcal{Z}}\right) = \boldsymbol{\Lambda}\left(\mathbf{I} - \mathbf{P}_{\mathcal{Z}}\right)$. Thus $\widehat{\mathrm{AV}}\left(\hat{\mathcal{X}}^{\mathrm{CGR}}_{3d}\right) = \mathcal{X}'_{3d}\boldsymbol{\Lambda}\left(\mathbf{I} - \mathbf{P}_{\mathcal{Z}}\right)\mathcal{X}_{3d}$. Now, composite calibration at domain level involves the design matrix $\mathcal{Z}_d = (\mathcal{X}_d,\mathbf{D})$; no need to restrict $\mathbf{D}$ to the domain U_d. The resulting CGR estimator is $\check{\mathcal{X}}^{\mathrm{CGR}}_{3d} = \hat{\mathcal{X}}_{3d} + \check{\mathcal{B}}_d\left(\mathbf{t}_{\mathcal{Z}_d} - \hat{\mathcal{Z}}_d\right)$ where $\check{\mathcal{B}}_d = \mathcal{X}'_{3d}\boldsymbol{\Lambda}\mathcal{Z}_d\left(\mathcal{Z}'_d\boldsymbol{\Lambda}\mathcal{Z}_d\right)^{-1}$. As with $\hat{\mathcal{X}}^{\mathrm{CGR}}_{3d}$ above, it can be shown that $\widehat{\mathrm{AV}}\left(\check{\mathcal{X}}^{\mathrm{CGR}}_{3d}\right) = \mathcal{X}'_{3d}\boldsymbol{\Lambda}\left(\mathbf{I} - \mathbf{P}_{\mathcal{Z}_d}\right)\mathcal{X}_{3d}$, where $\mathbf{P}_{\mathcal{Z}_d} = \mathcal{Z}_d\left(\mathcal{Z}'_d\boldsymbol{\Lambda}\mathcal{Z}\right)^{-1}_d\mathcal{Z}'_d\boldsymbol{\Lambda}$. Then $\widehat{\mathrm{AV}}\left(\hat{\mathcal{X}}^{\mathrm{CGR}}_{3d}\right) -$ $\widehat{\mathrm{AV}}\left(\check{\mathcal{X}}^{\mathrm{CGR}}_{3d}\right) = \mathcal{X}'_{3d}\boldsymbol{\Lambda}\left(\mathbf{P}_{\mathcal{Z}_d} - \mathbf{P}_{\mathcal{Z}}\right)\mathcal{X}_{3d}$. Noticing that $\mathcal{X}'_{3d}\boldsymbol{\Lambda}\mathcal{Z} = \mathcal{X}'_{3d}\boldsymbol{\Lambda}\mathcal{Z}_d$, we can write $\mathbf{P}_{\mathcal{Z}} =$ $\mathcal{Z}_d\left(\mathcal{Z}'\boldsymbol{\Lambda}\mathcal{Z}\right)^{-1}\mathcal{Z}'_d\boldsymbol{\Lambda}$. It is trivial then to show that $\left(\mathbf{P}_{\mathcal{Z}_d} - \mathbf{P}_{\mathcal{Z}}\right) = \left(\mathbf{P}_{\mathcal{Z}_d} - \mathbf{P}_{\mathcal{Z}}\right)^2$, and since the matrix $\boldsymbol{\Lambda}$ is diagonal with positive entries, it follows that $\mathcal{X}'_{3d}\boldsymbol{\Lambda}\left(\mathbf{P}_{\mathcal{Z}_d} - \mathbf{P}_{\mathcal{Z}}\right)\mathcal{X}_{3d} > \mathbf{0}$ and hence $\widehat{\mathrm{AV}}\left(\check{\mathcal{X}}^{\mathrm{CGR}}_{3d}\right) < \widehat{\mathrm{AV}}\left(\hat{\mathcal{X}}^{\mathrm{CGR}}_{3d}\right)$.

Under the conditions of part (b), $\boldsymbol{\Lambda} = \boldsymbol{\Lambda}^0$ and the CGR domain estimator is identical to the COR domain estimator $\hat{\mathcal{X}}^{\mathrm{COR}}_{3d} = \hat{\mathcal{X}}_{3d} - \hat{\mathcal{B}}^0_d\hat{\mathcal{X}}$, where $\hat{\mathcal{B}}^0_d = \mathcal{X}'_{3d}\boldsymbol{\Lambda}^0\mathcal{X}\left(\mathcal{X}'\boldsymbol{\Lambda}^0\mathcal{X}\right)^{-1}$. The associated matrix of regression residuals is $(\mathbf{I} - \mathbf{P}_{\mathcal{X}})\mathcal{X}_{3d}$, with $\mathbf{P}_{\mathcal{X}} = \mathcal{X}\left(\mathcal{X}'\boldsymbol{\Lambda}^0\mathcal{X}\right)^{-1}\mathcal{X}'\boldsymbol{\Lambda}^0$. Then $\widehat{\mathrm{AV}}\left(\hat{\mathcal{X}}^{\mathrm{COR}}_{3d}\right) =$ $\mathcal{X}'_{3d}\left(\mathbf{I} - \mathbf{P}_{\mathcal{X}}\right)'\boldsymbol{\Lambda}^0\left(\mathbf{I} - \mathbf{P}_{\mathcal{X}}\right)\mathcal{X}_{3d} = \mathcal{X}'_{3d}\boldsymbol{\Lambda}^0\left(\mathbf{I} - \mathbf{P}_{\mathcal{X}}\right)\mathcal{X}_{3d}$. On the other hand, for the estimator $\check{\mathcal{X}}^{\mathrm{COR}}_{3d} = \check{\mathcal{X}}_{3d} - \check{\mathcal{B}}^0_d\hat{\mathcal{X}}$, where $\check{\mathcal{B}}^0_d = \mathcal{X}'_{3d}\boldsymbol{\Lambda}^0\mathcal{X}_d\left(\mathcal{X}'_d\boldsymbol{\Lambda}^0\mathcal{X}_d\right)^{-1}$ we have $\widehat{\mathrm{AV}}\left(\check{\mathcal{X}}^{\mathrm{COR}}_{3d}\right) = \mathcal{X}'_{3d}\boldsymbol{\Lambda}^0$ $\left(\mathbf{I} - \mathbf{P}_{\mathcal{X}_d}\right)\mathcal{X}_{3d}$, with $\mathbf{P}_{\mathcal{X}_d} = \mathcal{X}_d\left(\mathcal{X}'_d\boldsymbol{\Lambda}^0\mathcal{X}_d\right)^{-1}\mathcal{X}'_d\boldsymbol{\Lambda}^0$. Then $\widehat{\mathrm{AV}}(\hat{\mathcal{X}}^{\mathrm{COR}}_{3d}) - \widehat{\mathrm{AV}}(\check{\mathcal{X}}^{\mathrm{COR}}_{3d}) = \mathcal{X}'_{3d}\boldsymbol{\Lambda}^0(\mathbf{P}_{\mathcal{X}_d} -$ $\mathbf{P}_{\mathcal{X}})\mathcal{X}_{3d}$. Notice that $\mathcal{X}'_{3d}\boldsymbol{\Lambda}^0\mathcal{X}_d = \mathcal{X}'_{3d}\boldsymbol{\Lambda}^0\mathcal{X}_{3d}$ and since $\boldsymbol{\Lambda}^0$ is diagonal $\mathcal{X}'_{3d}\boldsymbol{\Lambda}^0\mathcal{X} = \mathcal{X}'_{3d}\boldsymbol{\Lambda}^0\mathcal{X}_{3d}$. It follows that $\mathcal{X}'_{3d}\boldsymbol{\Lambda}^0\left(\mathbf{P}_{\mathcal{X}_d} - \mathbf{P}_{\mathcal{X}}\right)\mathcal{X}_{3d} = \mathcal{X}'_{3d}\boldsymbol{\Lambda}^0\left(\mathbf{P}_{\mathcal{X}_d} - \mathbf{P}_{\mathcal{X}}\right)^2\mathcal{X}_{3d}$ and hence $\widehat{\mathrm{AV}}\left(\check{\mathcal{X}}^{\mathrm{COR}}_{3d}\right) < \widehat{\mathrm{AV}}\left(\hat{\mathcal{X}}^{\mathrm{COR}}_{3d}\right)$.

For parts (a') and (b'), the proof is the same as in (a) and (b), in view of the proof of Theorem 1.

References

Andersson, P.G., and Thorburn, D. (2005). An optimal calibration distance leading to the optimal regression estimator. *Survey Methodology*, 31, 1, 95-99.

Australian Bureau of Statistics (2011). Household Expenditure Survey and Survey of Income and Housing, User Guide, Australia, 2009-10 (cat. no. 6503.0).

Chipperfield, J.O., and Steel, D.G. (2009). Design and estimation for split questionnaire surveys. *Journal of Official Statistics,* 25, 227-244.

Chipperfield, J.O., and Steel, D.G. (2011). Efficiency of split questionnaire surveys. *Journal of Statistical Planning and Inference,* 141, 1925-1932.

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

Fuller, W.A. (1990). Analysis of repeated surveys. *Survey Methodology,* 16, 2, 167-180.

Gonzalez, J.M., and Eltinge, J.L. (2007). Multiple matrix sampling: A review. *Proceedings of the Survey Research Methods Section,* American Statistical Association, 3069-3075.

Gonzalez, J.M., and Eltinge, J.L. (2008). Adaptive matrix sampling for the consumer expenditure quarterly interview survey. *Proceedings of the Survey Research Methods Section,* American Statistical Association, 3069-3075.

Hidiroglou, M.A. (2001). Double sampling. *Survey Methodology,* 27, 2, 143-154.

Houbiers, M. (2004). Towards a social statistical database on unified estimates at Statistics Netherlands. *Journal of Official Statistics,* 20, 55-75.

Jones, R.G. (1980). Best linear unbiased estimators for repeated surveys. *Journal of the Royal Statistical Society, Serie B,* 42, 221-226.

Kim, J.K., and Rao, J.N.K. (2012). Combining data from two independent surveys: A model-assisted approach. *Biometrika,* 99, 1, 85-100.

Merkouris, T. (2004). Combining independent regression estimators from multiple surveys. *Journal of the American Statistical Association,* 99, 1131-1139.

Merkouris, T. (2010). Combining information from multiple surveys by using regression for more efficient small domain estimation. *Journal of the Royal Statistical Society, Serie B,* 72, 27-48.

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

Montanari, G.E. (1998). On regression estimation of finite population means. *Survey Methodology,* 24, 1, 69-77.

Raghunathan, T.E., and Grizzle, J.E. (1995). A split questionnaire survey design. *Journal of the American Statistical Association,* 90, 54-63.

Rao, J.N.K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics*, 10, 153-165.

Renssen, R.H. (1998). Use of statistical matching techniques in calibration estimation. *Survey Methodology*, 24, 2, 171-183.

Renssen, R.H., and Nieuwenbroek, N.J. (1997). Aligning estimates for common variables in two or more sample surveys. *Journal of the American Statistical Association,* 92, 368-375.

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

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

Smith, P. (2009). Survey harmonization in official household surveys in the United Kingdom. *Proceedings of the ISI World Statistical Congresses*, Dublin.

Thomas, N., Raghunathan, T.E., Schenker, N., Katzoff, M.J. and Johnson, C.L. (2006). An evaluation of matrix sampling methods using data from the National Health and Nutrition Examination Survey. *Survey Methodology*, 32, 2, 217-231.

Wolter, K.M. (1979). Composite estimation in finite populations. *Journal of the American Statistical Association*, 74, 604-613.

Wu, C. (2004). Combining information from multiple surveys through the empirical likelihood method. *Canadian Journal of Statistics*, 32, 15-26.

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 30, No. 4, 2014

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. 1, 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. 1, March/mars 2015

CONTENTS TABLE DES MATIÈRES

Volume 43, No. 2, June/juin 2015

GUIDELINES FOR MANUSCRIPTS

Authors are invited to submit their articles in English or French in electronic form to the Editor, (smj@statcan.gc.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.