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

Survey Methodology

Survey Methodology
44-2

SURVEY
METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 2005
•
VOLUME 31
•
NUMBER 1

Release date: December 20, 2018

Statistics Canada Statistique Canada

Canada

How to obtain more information

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

You can also contact us by

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

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

• Statistical Information Service	1-800-263-1136
• National telecommunications device for the hearing impaired	1-800-363-7629
• Fax line	1-514-283-9350

Depository Services Program

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

Standards of service to the public

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

Note of appreciation

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

Standard table symbols

The following symbols are used in Statistics Canada publications:

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

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology

Catalogue No. 12-001-XPB

A journal published by Statistics Canada

December 2018

•

Volume 44

•

Number 2

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, (statcan.smj-rte.statcan@canada.ca, Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6). For formatting instructions, please see the guidelines provided in the journal and on the web site (www.statcan.gc.ca/surveymethodology).

Survey Methodology

A Journal Published by Statistics Canada

Volume 44, Number 2, December 2018

Special edition

Contents

Invited papers

Vol. 44, No. 2, p. 149
Statistics Canada, Catalogue No. 12-001-X

In this issue

Dear readers,

We are pleased to be the co-editors of this special issue of *Survey Methodology*. It contains 10 articles selected from all the presentations given at the 9th Colloque francophone sur les sondages, held in Gatineau from October 11 to 14, 2016.

The first three articles of this issue discuss various aspects of small area estimation. The article by Rao, Rubin-Bleuer and Estevao proposes an estimator of the design mean square error and studies its properties. In their article, Bertarelli, Ranalli, Bartolucci, D'Alo and Solari consider a latent Markov model to estimate the number of employed and unemployed people for various small areas and apply their model to data from the Italian Labour Force Survey. Finally, the article by De Moliner and Goga compares four methods for estimating mean electricity consumption curves for small areas.

The next three articles deal with sampling problems. The article by Grafström and Matei introduces sample coordination procedures for spatially balanced sampling designs. The article by Ida, Rivest and Daigle reviews two balanced sampling methods and compares them by means of a simulation study. Rebecq and Merly-Alpa study the problem of sample allocation for stratified sampling designs with simple random sampling in each stratum. The authors propose a compromise between optimal allocation and proportional allocation that leads to weakly dispersed weights.

The last four articles in this issue examine different aspects of survey sampling methods. The article by Juillard and Chauvet studies the problem of point and variance estimation in the presence of unit non-response in panel surveys. In their article, Bosa, Godbout, Mills and Picard propose a decomposition of the variance in the presence of imputation, which is used to quantify the effect of converting a non-respondent to a respondent. They also evaluate their method through a simulation. Deroyon and Favre-Martinoz extend two methods for determining the winsorization threshold to the case of Poisson sampling designs and compare~~s~~ them empirically. Finally, the article by Tirari and Hdioud proposes a weighting effect to quantify the impact of calibration on accuracy using an approach based on the design and the model.

We hope you enjoy this issue!

Jean-Francois Beaumont and David Haziza[1].
Guest co-editors of this special issue

[1] Jean-François Beaumont, International Cooperation and Methodology Innovation Centre, Statistics Canada, R.H. Coats Bldg, 25th Floor, 100 Tunney's Pasture Driveway, Ottawa, Ontario, K1A 0T6. E-mail: jean-francois.beaumont@canada.ca; David Haziza, Département de mathématiques et statistique, Université de Montréal, Montréal, Canada, H3C 3J7. E-mail: haziza@dms.umontreal.ca.

Survey Methodology, December 2018
Vol. 44, No. 2, pp. 151-166
Statistics Canada, Catalogue No. 12-001-X

Measuring uncertainty associated with model-based small area estimators

J.N.K. Rao, Susana Rubin-Bleuer and Victor M. Estevao[1]

Abstract

Domains (or subpopulations) with small sample sizes are called small areas. Traditional direct estimators for small areas do not provide adequate precision because the area-specific sample sizes are small. On the other hand, demand for reliable small area statistics has greatly increased. Model-based indirect estimators of small area means or totals are currently used to address difficulties with direct estimation. These estimators are based on linking models that borrow information across areas to increase the efficiency. In particular, empirical best (EB) estimators under area level and unit level linear regression models with random small area effects have received a lot of attention in the literature. Model mean squared error (MSE) of EB estimators is often used to measure the variability of the estimators. Linearization-based estimators of model MSE as well as jackknife and bootstrap estimators are widely used. On the other hand, National Statistical Agencies are often interested in estimating the design MSE of EB estimators in line with traditional design MSE estimators associated with direct estimators for large areas with adequate sample sizes. Estimators of design MSE of EB estimators can be obtained for area level models but they tend to be unstable when the area sample size is small. Composite MSE estimators are proposed in this paper and they are obtained by taking a weighted sum of the design MSE estimator and the model MSE estimator. Properties of the MSE estimators under the area level model are studied in terms of design bias, relative root mean squared error and coverage rate of confidence intervals. The case of a unit level model is also examined under simple random sampling within each area. Results of a simulation study show that the proposed composite MSE estimators provide a good compromise in estimating the design MSE.

Key Words: Area and unit level models; Composite estimators of design mean squared error; Empirical best linear unbiased predictor; Estimating design mean squared error.

1 Introduction

Sample survey data are often used to produce estimates of domain (subpopulation) totals or means. Traditional direct estimators for domains, including calibration estimators that use known population totals of auxiliary variables, are designed to provide reliable estimators for domains with large domain-specific sample sizes. However, direct estimators do not provide adequate precision for domains with small sample sizes (called small areas). Yet the demand for reliable small area statistics has greatly increased in recent years. It is therefore necessary to resort to indirect estimators that borrow information from related areas through known auxiliary information such as censuses and administrative records, to increase the efficiency. Indirect estimators based on explicit linking models are widely used; in particular, empirical best (EB) estimators based on area level or unit level linear regression models with random area effects. A detailed account of EB estimation under those models is given by Rao and Molina (2015), Chapters 6 and 7. Section 2 presents EB estimators of small area means under basic area level and unit level models.

EB-type model based estimators are often deemed suitable by National Statistical Agencies to produce official statistics, after careful external evaluations. For example, Beaumont and Bocci (2016) compared EB and direct estimates of unemployment rate for small areas obtained from the Canadian Labour Force Survey

[1] J.N.K. Rao, Distinguished Research Professor, Carleton University, Ottawa, Ontario. E-mail: jrao34@rogers.com; Susana Rubin-Bleuer, Adjunct Research Professor, Carleton University, Ottawa, Ontario; Victor M. Estevao, Senior Methodologist, Statistics Canada, R.H. Coats Bldg, 25th Floor, 100 Tunney’s Pasture, Ottawa, Ontario, K1A 0T6. E-mail: victor.estevao@canada.ca.

(LFS) to “gold standard” estimates obtained from the much larger National Household Survey (comparable to long form census) and found that the relative error of EB estimates is much smaller than the corresponding direct estimates. The authors used a basic area level linear regression model with random area effects to produce EB estimates. External evaluations were first used in the pioneering paper by Fay and Herriot (1979) under a basic area level model to produce estimates of mean income for small places in the United States.

Model mean squared error (MSE) of the EB estimators is often used to measure the variability of the estimators. In particular, linearization-based estimators of model MSE as well as jackknife and bootstrap estimators are widely used. Section 3 gives a brief account of model-based MSE estimation, including estimators based on unconditional and conditional frameworks.

The literature on estimating model MSE is very impressive, but National Statistical agencies are often interested in estimating the design MSE of EB estimators in line with the traditional design MSE estimators of direct estimators for large areas with adequate sample sizes (Pfeffermann and Gilboa, 2017). Estimators of design MSE of EB estimators for the basic area level model can be obtained but they tend to be unstable when the area sample size is small. To address this problem, Section 4 proposes composite MSE estimators obtained by taking a weighted sum of the design MSE estimator and the model MSE estimator. The case of unit level models is also studied under simple random sampling within areas. Section 5 reports the results of simulation studies on the performance of the proposed composite MSE estimators in terms of design absolute relative bias (ARB), relative root mean squared error (RRMSE) and coverage of confidence intervals. Both area level and unit level models are considered in the simulation study. Finally, some conclusions are given in Section 6.

2 EB estimators

In this section, we present EB estimators of small area means or totals, denoted by $\theta_i,$ for m areas with small sample sizes. For area level models we assume that direct estimators $\hat{\theta}_i$ and associated area level covariates $\mathbf{z}_i$ are available for the m areas, where $\mathbf{z}_i$ is a $p \times 1$ vector. In the case of unit level models, we assume that unit level data $\{(y_{ij}, \mathbf{x}_{ij}), \; j = 1, \ldots, n_i; \; i = 1, \ldots, m\}$ are available for the sampled areas, where n_i is the sample size in area i and $\mathbf{x}_{ij}$ is a $p \times 1$ vector of covariates that can include area level covariates. We assume that the area population means $\bar{\mathbf{X}}_i$ are known.

2.1 Basic area level model

We assume that the direct estimator $\hat{\theta}_i$ is design unbiased (either exactly or approximately for large overall sample size n). For example, estimators calibrated to known overall means of auxiliary variables are approximately unbiased. We can express this assumption as a sampling model $\hat{\theta}_i = \theta_i + e_i,$ where the sampling error e_i has zero mean and variance $\psi_i.$ We further assume that the sampling variance ψ_i is known and not random. In practice, the estimators of the sampling variances are smoothed and the resulting smoothed estimator is taken as a proxy for $\psi_i.$ Beaumont and Bocci (2016) propose a method of smoothing the sampling variances in the context of Canadian LFS. The model linking the areas assumes that the θ_i are random, obeying the "matching" linking model $\theta_i = \mathbf{z}_i'\boldsymbol{\beta} + v_i$, where the random area effect v_i has zero mean and variance σ_v^2 and is independent of the sampling error e_i. We further assume normality of v_i and e_i.

Combining the sampling model with the linking model leads to the basic area level model

$$\hat{\theta}_i = \mathbf{z}_i'\boldsymbol{\beta} + v_i + e_i, \quad v_i \overset{\text{iid}}{\sim} N(0, \sigma_v^2), \quad e_i \overset{\text{id}}{\sim} N(0, \psi_i), \quad i = 1, \ldots, m. \tag{2.1}$$

Main advantages of model (2.1) are that it takes account of the sampling design through the sampling model on the direct estimators and that it requires only area level covariates, which are more readily available than unit level covariates.

For known model parameters $(\boldsymbol{\beta}, \sigma_v^2)$, the "best" estimator of θ_i is given by

$$\tilde{\theta}_i^B = \tilde{E}\left(\theta_i \,|\, \hat{\theta}_i, \boldsymbol{\beta}, \sigma_v^2\right) = \gamma_i \hat{\theta}_i + (1 - \gamma_i)\, \mathbf{z}_i'\boldsymbol{\beta}, \tag{2.2}$$

where $\gamma_i = \sigma_v^2 / (\sigma_v^2 + \psi_i)$. The best estimator (2.2) is unbiased for θ_i in the sense that $E\left(\tilde{\theta}_i^B - \theta_i\right) = 0$, where the expectation is with respect to the assumed model (2.1), that is, design-model expectation (Rubin-Bleuer and Schiopu-Kratina, 2005). It follows from (2.2) that more weight is given to the direct estimator $\hat{\theta}_i$ if the model variance σ_v^2 is large relative to the sampling variance ψ_i, and more weight given to the synthetic estimator $\mathbf{z}_i'\boldsymbol{\beta}$ if the sampling variance is large.

The mean squared error (MSE) of the best estimator under the model (2.1) is given by

$$\text{MSE}\left(\tilde{\theta}_i^B\right) = E\left(\tilde{\theta}_i^B - \theta_i\right)^2 = \gamma_i \psi_i, \tag{2.3}$$

where the term $\gamma_i \psi_i$ is often denoted by $g_{1i}(\sigma_v^2)$. It follows from (2.3) that the optimal estimator leads to significant reduction in MSE over the direct estimator if γ_i is small or the model variance is relatively small compared to the total variance $\sigma_v^2 + \psi_i$. This result provides a convincing justification for using the model-based approach to produce small area estimates.

In practice, the model parameters are not known and we replace the parameters in (2.2) by restricted maximum likelihood (REML) estimators $\left(\hat{\boldsymbol{\beta}}, \hat{\sigma}_v^2\right)$ to get the empirical best (EB) estimator:

$$\hat{\theta}_i^{\text{EB}} = \hat{\gamma}_i \hat{\theta}_i + (1 - \hat{\gamma}_i)\, \mathbf{z}_i'\hat{\boldsymbol{\beta}}. \tag{2.4}$$

Rao and Molina (2015), Chapter 6, give details of REML estimation of the model parameters.

2.2 Basic unit level model

We now turn to a basic unit level model which uses unit level sample data $\{(y_{ij}, \mathbf{x}_{ij}),\ j = 1, \ldots, n_i;\ i = 1, \ldots, m\}$, where n_i is the sample size in area i. We assume that the area population means $\overline{\mathbf{X}}_i$ are known. We further assume a basic unit level nested error linear regression model for the population and the same model holds for the sample (Battese, Harter and Fuller, 1988). The sample model is given by

$$y_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + v_i + e_{ij}, \quad j = 1, \ldots, n_i; \ i = 1, \ldots, m, \tag{2.5}$$

where the area random effects $v_i \overset{\text{iid}}{\sim} N(0, \sigma_v^2)$ are assumed to be independent of the unit errors $e_{ij} \overset{\text{iid}}{\sim} N(0, \sigma_e^2)$. Unit level models can lead to significant gains in efficiency over area level models because the model parameters can be estimated more accurately using all the observations in the overall sample, unlike area level models.

For known parameters $(\boldsymbol{\beta}, \sigma_v^2, \sigma_e^2)$, the "best" estimator of the area mean $\bar{Y}_i$ is given by

$$\hat{\bar{Y}}_i^B = E\left[\bar{Y}_i \left|\left(y_{ij},\ j = 1, \ldots, n_i;\ \mathbf{x}_{ij},\ j = 1, \ldots, N_i\right), \boldsymbol{\beta}, \sigma_v^2, \sigma_e^2\right.\right] = \bar{\mathbf{X}}_i'\boldsymbol{\beta} + a_i\left(\bar{y}_i - \bar{\mathbf{x}}_i'\boldsymbol{\beta}\right), \tag{2.6}$$

where $\bar{y}_i$ and $\bar{\mathbf{x}}_i$ are the sample means, $a_i = (1 - f_i)\gamma_i + f_i$ with sampling fraction $f_i = n_i / N_i$ and $\gamma_i = \sigma_v^2 / (\sigma_v^2 + \sigma_e^2 / n_i)$, and N_i is the number of population units in area i (Rao and Molina, 2015, Chapter 7). If the area population size N_i is large and $f_i \approx 0$, then (2.6) reduces to a weighted combination of the "sample regression" estimator $\bar{y}_i + (\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_i)' \boldsymbol{\beta}$ and the regression synthetic estimator $\bar{\mathbf{X}}_i'\boldsymbol{\beta}$ with weights γ_i and $1 - \gamma_i$ respectively. We denote this approximation to $\hat{\bar{Y}}_i^B$ by $\hat{\mu}_i^B$. As the area sample size n_i increases, the optimal estimator gives more weight to the sample regression estimator. In practice, we replace the model parameters by REML estimators $\left(\hat{\boldsymbol{\beta}}, \hat{\sigma}_v^2, \hat{\sigma}_e^2\right)$ to get the EB estimator $\hat{\bar{Y}}_i^{\text{EB}}$ or $\hat{\mu}_i^{\text{EB}}$.

The EB estimator under the unit level model (2.5) does not account for the survey weights w_{ij}, unlike the area level model. As a result, the EB estimator is not design consistent as the area sample size increases, unless the weights are all equal within the area.

The MSE of $\hat{\mu}_i^B$ is equal to $g_{1i}(\sigma_v^2, \sigma_e^2) = \gamma_i (\sigma_e^2 / n_i)$ while the MSE of the sample regression estimator is equal to σ_e^2 / n_i. It now follows that the optimal estimator leads to significant reduction in MSE over the sample regression estimator if γ_i is small or the model variance σ_v^2 is small relative to the total variance $\sigma_v^2 + \sigma_e^2 / n_i$.

3 Model-based MSE estimators

In this section, we focus on the model-based MSE of EB estimators under the basic area level and unit level models. No closed form expressions for MSE exist, except for a few special cases. This problem has attracted much attention in the SAE literature, leading to second-order approximations to MSE which in turn are used to obtain second-order unbiased estimators of MSE under the assumed models.

3.1 Basic area-level model

We focus on REML estimators of model parameters, denoted $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}_v^2$. A second-order unbiased estimator of unconditional model MSE of the EB estimator is given by

$$\text{mse}\left(\hat{\theta}_i^{\text{EB}}\right) = g_{1i}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2) + 2g_{3i}(\hat{\sigma}_v^2). \tag{3.1}$$

Here the leading term in (3.1) is given by (2.3) with σ_v^2 replaced by $\hat{\sigma}_v^2$ and the remaining two terms in (3.1) are of lower order and account for the estimation of $\boldsymbol{\beta}$ and σ_v^2, respectively (see Rao and Molina, 2015, Chapter 6 for details). The MSE estimator (3.1) is positive and second-order unbiased in the sense that its bias is of lower order than $1/m$ for large m. Parametric bootstrap methods have also been used to obtain a MSE estimator. However, the resulting MSE estimator is not second-order unbiased and an additional bias adjustment is made to ensure second-order unbiasedness. Those adjustments typically require double bootstrap methods and some of the adjusted bootstrap MSE estimators may take negative values; see Rao and Molina (2015), Chapter 6.

3.2 Basic unit-level model

We again focus on REML estimation of model parameters in the unit level model (2.5). A positive second-order unbiased estimator of the unconditional MSE of the EB estimator $\hat{\mu}_i^{\text{EB}}$ is given by

$$\text{mse}\left(\hat{\mu}_i^{\text{EB}}\right) = g_{1i}\left(\hat{\sigma}_v^2, \hat{\sigma}_e^2\right) + g_{2i}\left(\hat{\sigma}_v^2, \hat{\sigma}_e^2\right) + 2g_{3i}\left(\hat{\sigma}_v^2, \hat{\sigma}_e^2\right), \tag{3.2}$$

where the first term is the leading term given in Section 2.2, the second term is due to estimating $\boldsymbol{\beta}$ and the last term is due to estimating σ_v^2 and σ_e^2. The EB estimator $\hat{\mu}_i^{\text{EB}}$ and the associated unconditional MSE estimator (3.2) are valid when the sampling fraction f_i is negligible. We refer the reader to (Rao and Molina, 2015, Section 7.2.3) for MSE estimation in the case of non-negligible sampling fractions.

4 Design MSE estimation

In this section we first study design MSE estimation and then propose composite MSE estimation that provides a balance between the design bias and the coefficient of variation.

4.1 Area-level model

We now turn to estimating the design MSE of the EB estimator by treating the small area parameters θ_i as fixed unknown parameters. As noted in the introduction, survey statisticians are often interested in estimating the design MSE of EB estimators in line with the traditional design MSE estimators of direct estimators for large areas with adequate sample sizes. The design MSE is given by $\text{MSE}_d\left(\hat{\theta}_i^{\text{EB}}\right) = E\left[\left(\hat{\theta}_i^{\text{EB}} - \theta_i\right)^2 \middle| \boldsymbol{\theta}\right]$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)'$ is the vector of area means.

Expressing $\hat{\theta}_i^{\text{EB}}$ as $\hat{\theta}_i + h_i\left(\hat{\boldsymbol{\theta}}\right)$ with $h_i\left(\hat{\boldsymbol{\theta}}\right) = -(1-\hat{\gamma}_i)\left(\hat{\theta}_i - \mathbf{z}_i'\hat{\boldsymbol{\beta}}\right)$, an exactly unbiased estimator of the design MSE is given by

$$\text{mse}_d\left(\hat{\theta}_i^{\text{EB}}\right) = \psi_i + 2\psi_i\left[\partial h_i\left(\hat{\boldsymbol{\theta}}\right)/\partial\hat{\theta}_i\right] + h_i^2\left(\hat{\boldsymbol{\theta}}\right). \tag{4.1}$$

Datta, Kubokawa, Molina and Rao (2011) give an explicit expression for the derivative in the second term of (4.1) in the case of REML estimators of model parameters. The estimator (4.1) can take negative values and can be very unstable in terms of relative root mean squared error (RRMSE) as shown by Datta et al. (2011). It follows that (4.1) is not a reliable estimator of the design MSE, although it is design unbiased.

Our simulation results in Section 5 study the conditional properties of the MSE estimators (3.1) and (4.1) in the design-based framework.

Some theoretical insights can be obtained by focusing on the case of known model parameters and considering the best estimator (2.2) of the area mean θ_i. In this case, Rivest and Belmonte (2000) obtained a design-unbiased estimator given by

$$\operatorname{mse}_d\left(\tilde{\theta}_i^B\right) = \gamma_i \psi_i + (1 - \gamma_i)^2 \left[\left(\hat{\theta}_i - \mathbf{z}_i'\boldsymbol{\beta}\right)^2 - (\psi_i + \sigma_v^2)\right]. \tag{4.2}$$

Note that for a large sampling variance ψ_i we have $\gamma_i \approx 0$ and (4.2) reduces to

$$\operatorname{mse}_d\left(\tilde{\theta}_i^B\right) \approx \left(\hat{\theta}_i - \mathbf{z}_i'\boldsymbol{\beta}\right)^2 - \psi_i. \tag{4.3}$$

It follows from (4.3) that the MSE estimator can take negative values and, in fact, the probability of getting a negative value is close to 0.5 when γ_i is close to zero or sampling variance ψ_i is large. In this special case of known model parameters, we can study the design bias of the model MSE estimator of (2.2), given by $\operatorname{mse}\left(\tilde{\theta}_i^B\right) = \gamma_i \psi_i$, when averaged over the areas. It can be shown that the average design bias converges in model probability to zero as $m \to \infty$ (Rao and Molina, 2015, page 287). This result suggests that the model MSE estimator should perform well in terms of average design bias, provided the assumed model is valid.

The design-unbiased estimator (4.1) is not usable in practice when it takes a negative value for the sample at hand. Therefore, we propose a modification of (4.1) that leads to a positive MSE estimator. We denote the modified MSE estimator by $\text{mod-mse}_d\left(\hat{\theta}_i^{\text{EB}}\right)$. It uses (4.1) when it takes a positive value for the sample at hand and replaces (4.1) by the model MSE estimate (3.1) when (4.1) takes a negative value. It is possible to use some other positive MSE estimate, for example a naïve positive design-based MSE estimator proposed by Pfeffermann and Gilboa (2017). We have not studied this modification in our simulation study.

We now propose composite estimators of the design MSE that attempt to provide a balance between design bias and RRMSE. One composite estimator is obtained by taking a weighted average of the design MSE estimator (4.1) and the unconditional model MSE estimator (3.1) with weights $\hat{\gamma}_i$ and $(1 - \hat{\gamma}_i)$ respectively. This composite MSE estimator may be written as

$$\operatorname{mse}_{c1}\left(\hat{\theta}_i^{\text{EB}}\right) = \hat{\gamma}_i \operatorname{mse}_d\left(\hat{\theta}_i^{\text{EB}}\right) + (1 - \hat{\gamma}_i)\operatorname{mse}\left(\hat{\theta}_i^{\text{EB}}\right). \tag{4.4}$$

It follows from (4.4) that less weight is given to the design MSE estimator when the sampling variance is large and this controls the RRMSE of the composite MSE estimator. Also, the composite MSE estimator has always a smaller design bias than the model MSE estimator. When $\hat{\gamma}_i$ (or the area sample size) is very small, another choice of the compositing weights is to replace $\hat{\gamma}_i$ by $\sqrt{\hat{\gamma}_i}$ and $1 - \hat{\gamma}_i$ by $1 - \sqrt{\hat{\gamma}_i}$ in (4.4). The resulting composite MSE estimator

$$\operatorname{mse}_{c2}\left(\hat{\theta}_i^{\text{EB}}\right) = \sqrt{\hat{\gamma}_i}\operatorname{mse}_d\left(\hat{\theta}_i^{\text{EB}}\right) + \left(1 - \sqrt{\hat{\gamma}_i}\right)\operatorname{mse}\left(\hat{\theta}_i^{\text{EB}}\right) \tag{4.5}$$

gives more weight to $\mathrm{mse}_d\left(\hat{\theta}_i^{\mathrm{EB}}\right)$ than (4.4) and thus performs better in terms of design bias at the expense of increased MSE. Similar to (4.4), the alternative composite MSE estimator (4.5) has always a smaller design bias than the model MSE estimator. Both (4.4) and (4.5) can also take on negative values but likely not as often due to their construction. To ensure positive composite MSE estimators, we make a modification similar to $\text{mod-mse}_d\left(\hat{\theta}_i^{\mathrm{EB}}\right)$ and replace (4.4) and (4.5) by the model MSE estimate (3.1) when they take negative values for the sample at hand. We denote the modified estimators by $\text{mod-mse}_{c1}\left(\hat{\theta}_i^{\mathrm{EB}}\right)$ and $\text{mod-mse}_{c2}\left(\hat{\theta}_i^{\mathrm{EB}}\right)$ respectively. In Section 5, we look at the performance of the two modified composite MSE estimators relative to the model MSE estimator (3.1) and the modified design MSE estimator in terms of ARB, RRMSE and coverage rate of confidence intervals.

4.2 Unit-level model

We focus on simple random sampling (SRS) without replacement in each area. Even for this special design, no closed form expressions for the design MSE of the EB estimator $\hat{\bar{Y}}_i^{\mathrm{EB}}$ and its estimator are available in the literature, unlike in the case of the area-level model. Therefore, we propose a heuristic method by evaluating the design MSE of the best estimator $\hat{\bar{Y}}_i^{B}$ given by (2.6), under SRS assuming all the model parameters are known and then estimating the design MSE. The resulting design-unbiased MSE estimator of the best estimator depends on the model parameters and we replace the model parameters by their REML estimators. The resulting MSE estimator is not design-unbiased for the design MSE of the EB estimator and it is likely to underestimate the true design MSE because the variability associated with the estimated model parameters is not taken into account. We study its design performance in a simulation study.

Under SRS without replacement within area i, we have

$$\hat{\bar{Y}}_i^{B} - \bar{Y}_i = a_i\left(\bar{u}_i - \bar{U}_i\right) - (1 - a_i)\bar{U}_i, \tag{4.6}$$

where $\bar{u}_i$ is the area sample mean and $\bar{U}_i$ is the area population mean of the values $u_{ij} = y_{ij} - \mathbf{x}'_{ij}\boldsymbol{\beta}$. It follows from (4.6) that the design MSE of the best estimator is given by

$$\mathrm{MSE}_d\left(\hat{\bar{Y}}_i^{B}\right) = E_d\left(\hat{\bar{Y}}_i^{B} - \bar{Y}_i\right)^2 = a_i^2 V_d\left(\bar{u}_i\right) + (1 - a_i)^2 \bar{U}_i^2, \tag{4.7}$$

where

$$V_d\left(\bar{u}_i\right) = n_i^{-1}(1 - f_i)S_{ui}^2, \quad \text{and} \quad S_{ui}^2 = (N_i - 1)^{-1}\sum_{j=1}^{N_i}(u_{ij} - \bar{U}_i)^2, \tag{4.8}$$

noting that the cross-product term is zero under SRS.

It now follows from (4.7) and (4.8) that a design unbiased MSE estimator of the best estimator is given by

$$\mathrm{mse}_d\left(\hat{\bar{Y}}_i^{B}\right) = a_i^2 n_i^{-1}(1 - f_i)s_{ui}^2 + (1 - a_i)^2 \hat{\bar{U}}_i^{2D}, \tag{4.9}$$

where $\hat{\hat{U}}_i^{2D} = n_i^{-1}\sum_{j=1}^{n_i} u_{ij}^2 - N_i^{-1}(N_i - 1)\, s_{ui}^2$ and $s_{ui}^2 = (n_i - 1)^{-1}\sum_{j=1}^{n_i}\left(u_{ij} - \bar{u}_i\right)^2$. By replacing the model parameters in (4.9) by their REML estimators, a design-based MSE estimator of the EB estimator is obtained, denoted by $\text{mse}_d^*\left(\hat{\hat{Y}}_i^{\text{EB}}\right)$. This MSE estimator is likely to underestimate the design MSE of the EB estimator because the best estimator (2.6) does not account for the variability in the estimators of model parameters.

A composite MSE estimator, $\text{mse}_c^*\left(\hat{\hat{Y}}_i^{\text{EB}}\right)$, is now obtained by taking a weighted combination of $\text{mse}_d^*\left(\hat{\hat{Y}}_i^{\text{EB}}\right)$ and the model-based MSE estimator $\text{mse}\left(\hat{\hat{Y}}_i^{\text{EB}}\right)$ with weights $\hat{\gamma}_i$ and $1 - \hat{\gamma}_i$ respectively. It is given by

$$\text{mse}_c^*\left(\hat{\hat{Y}}_i^{\text{EB}}\right) = \hat{\gamma}_i\ \text{mse}_d^*\left(\hat{\hat{Y}}_i^{\text{EB}}\right) + (1 - \hat{\gamma}_i)\,\text{mse}\left(\hat{\hat{Y}}_i^{\text{EB}}\right). \tag{4.10}$$

Molina and Kominiak (2017) proposed parametric and non-parametric bootstrap estimators of the design MSE of $\hat{\hat{Y}}_i^{\text{EB}}$. They also obtained a composite MSE estimator, similar to (4.10), by using the non-parametric bootstrap (NPB) MSE estimator and the parametric bootstrap (PB) MSE estimator as the two components of the composite MSE estimator associated with $\hat{\gamma}_i$ and $(1 - \hat{\gamma}_i)$ respectively. As noted by the authors, a drawback with this composite MSE estimator is "that it requires to run both PB and NPB procedures for each area, which makes it computationally slower." Molina and Kominiak (2017) also proposed a parametric design bootstrap (PDB) composite MSE estimator. The PDB estimator avoids running both PB and NPB procedures for each area. Both bootstrap composite MSE estimators performed well in a design-based simulation study.

5 Simulation study

In this section, we report the results of limited simulation studies on the design performance of the proposed composite MSE estimators. Section 5.1 gives results for the area level model, and the unit level model results are reported in Section 5.2.

5.1 Area-level model

Following the simulation set up used by Datta et al. (2011), we employ model (2.1) with $m = 30$ areas, $\mathbf{z}_i = (1, z_{i1})'$ for $i = 1, \ldots, m,$ where the covariate values $z_{i1}, \ldots, z_{im}$ are generated independently from $N(-1, 1)$ and held fixed over the simulation runs. Further, $\boldsymbol{\beta} = (1, 1)'$, $\sigma_v^2 = 1$ and the sampling variance values are (2.0, 0.6, 0.5, 0.4, 0.2), with each different value of ψ_i assigned to six consecutive areas. Noting that $v_i \sim N(0, 1)$, we generate $\{\theta_i;\, i = 1, \ldots, m\}$ from the linking model $\theta_i = \mathbf{z}_i'\boldsymbol{\beta} + v_i$ and hold them fixed over the simulations to reflect the design-based approach conditioning on the area means θ_i. Then, $R = 100{,}000$ simulated samples $\left\{\hat{\theta}_i^{(r)}\!:\ i = 1, \ldots, m\right\}$, $r = 1, \ldots, R$ are generated from the sampling model $\hat{\theta}_i = \theta_i + e_i$ with the sampling error e_i generated from $N(0, \psi_i)$ for specified sampling variance ψ_i which is assumed fixed and known. We note that our simulation setup is not exactly design-based but it is "close enough" for the purposes of our study.

From the simulated data $\left\{\left(\hat{\theta}_i^{(r)},\, \mathbf{z}_i\right):\ i = 1, \ldots, m\right\}$ the EB estimates $\hat{\theta}_i^{\mathrm{EB}(r)}$ are computed and the MSE of $\hat{\theta}_i^{\mathrm{EB}}$ is approximated by

$$\mathrm{MSE}_i^{\mathrm{EB}} = R^{-1}\sum_{r=1}^{R}\left(\hat{\theta}_i^{\mathrm{EB}(r)} - \theta_i\right)^2. \tag{5.1}$$

The MSE estimators for each simulated sample are computed and averaged over the 100,000 simulation runs. We denote the means of the MSE estimators over the simulations as $\mathrm{mse}_i^{\mathrm{EB}}$, $\mathrm{mse}_{di}^{\mathrm{EB}}$, $\text{mod-mse}_{di}^{\mathrm{EB}}$, $\text{mod-mse}_{c1i}^{\mathrm{EB}}$ and $\text{mod-mse}_{c2i}^{\mathrm{EB}}$, corresponding to model, design unbiased, modified design unbiased, modified composite 1 and modified composite 2 MSE estimators, respectively. The relative bias (RB) of $\mathrm{mse}_i^{\mathrm{EB}}$ is given by

$$\mathrm{RB}_i^{\mathrm{EB}} = \left(\mathrm{mse}_i^{\mathrm{EB}} - \mathrm{MSE}_i^{\mathrm{EB}}\right)/\mathrm{MSE}_i^{\mathrm{EB}} \tag{5.2}$$

where $\mathrm{MSE}_i^{\mathrm{EB}}$ is given by (5.1). The absolute relative bias (ARB) is simply defined as $\mathrm{ARB}_i^{\mathrm{EB}} = \left|\,\mathrm{RB}_i^{\mathrm{EB}}\,\right|$. The terms $\mathrm{ARB}_{di}^{\mathrm{EB}}$, $\mathrm{ARB}_{di-\mathrm{mod}}^{\mathrm{EB}}$, $\mathrm{ARB}_{c1i-\mathrm{mod}}^{\mathrm{EB}}$ and $\mathrm{ARB}_{c2i-\mathrm{mod}}^{\mathrm{EB}}$ are defined in a similar manner.

We also compute the relative root mean squared error (RRMSE) of the MSE estimators over the simulations. We denote those values as $\mathrm{RRMSE}_i^{\mathrm{EB}}$, $\mathrm{RRMSE}_{di}^{\mathrm{EB}}$, $\mathrm{RRMSE}_{di-\mathrm{mod}}^{\mathrm{EB}}$, $\mathrm{RRMSE}_{c1i-\mathrm{mod}}^{\mathrm{EB}}$ and $\mathrm{RRMSE}_{c2i-\mathrm{mod}}^{\mathrm{EB}}$ for the model, design unbiased, modified design unbiased, modified composite 1 and modified composite 2 MSE estimators, respectively. Here RRMSE of the model MSE estimator is defined as

$$\mathrm{RRMSE}_i^{\mathrm{EB}} = \left\{R^{-1}\sum_{r=1}^{R}\left(\mathrm{mse}_i^{\mathrm{EB}(r)} - \mathrm{MSE}_i^{\mathrm{EB}}\right)^2\right\}^{1/2} \Big/ \mathrm{MSE}_i^{\mathrm{EB}}. \tag{5.3}$$

The RRMSE of the other MSE estimators are similarly defined.

We first compare the average over all the areas of $\mathrm{mse}_i^{\mathrm{EB}}$ to the average over all areas of $\mathrm{mse}_{di}^{\mathrm{EB}}$. We obtain 0.42 and 0.35 respectively, showing that the average of the model MSE estimator, 0.42, is close enough to the average of the design MSEs of the EB estimators, 0.35, confirming the theoretical result mentioned in Section 4.1. The theoretical result assumes known model parameters, while the simulation deals with the general case of unknown model parameters.

We next examine the probability of getting a negative value for the three MSE estimators: design unbiased, composite 1 and composite 2. Figure 5.1 shows the percentage of negative values over the simulations for each of the thirty areas. It is clear from Figure 5.1 that the probability of getting a negative value for the design unbiased MSE estimator can be as large as 50% for the first six areas (group 1) with much larger sampling variance relative to the remaining areas (group 2). On the other hand, it is negligible for the areas in group 2. The average probability over areas in group 1 is 45.67% compared to 0.03% in group 2. The probability of getting a negative value for the composite 1 MSE estimator is zero across all thirty areas, while the average probability for the composite 2 MSE estimator is 9.15% over areas in group 1 and zero over areas in group 2. The above results suggest that the composite 1 MSE estimator may not need modification even for areas with large sampling variances. Note that in the current simulation study the composite 1 and modified composite 1 MSE estimators are identical because no zero values were found for the composite 1 MSE estimator.

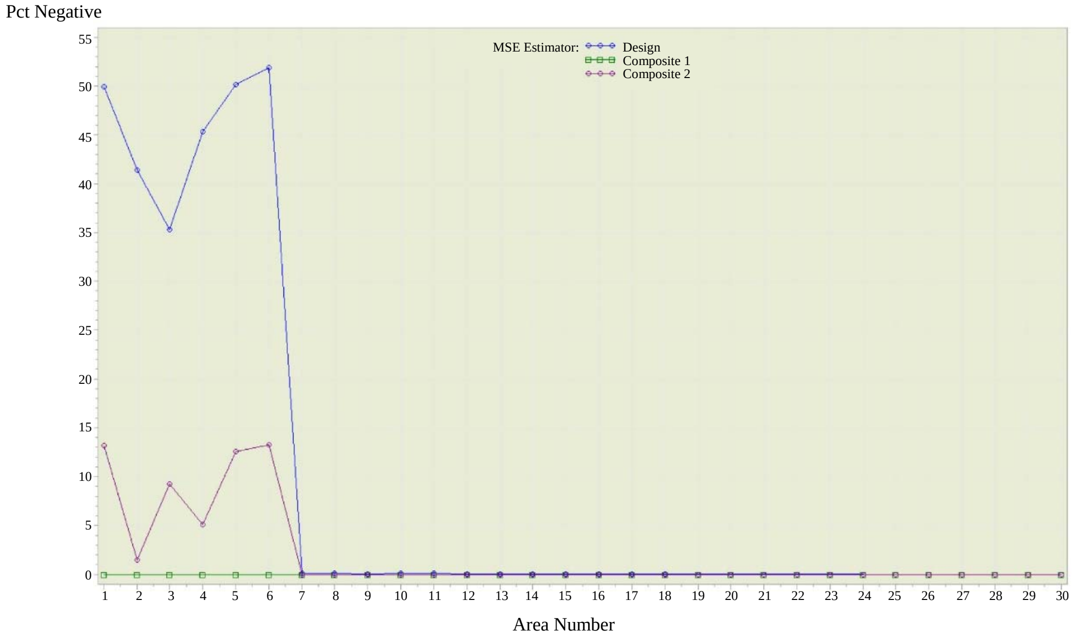


Figure 5.1 Plot of the percent of negative values of the MSE estimators: area level model.

We now turn to the ARB of the MSE estimators. Figure 5.2 shows the ARB values across all the thirty areas for the MSE estimators: model, design unbiased, modified design unbiased, modified composite 1 and modified composite 2 MSE estimators. Table 5.1 gives the mean % design ARB values as well as the mean % design RRMSE over the areas in group 1 and group 2.

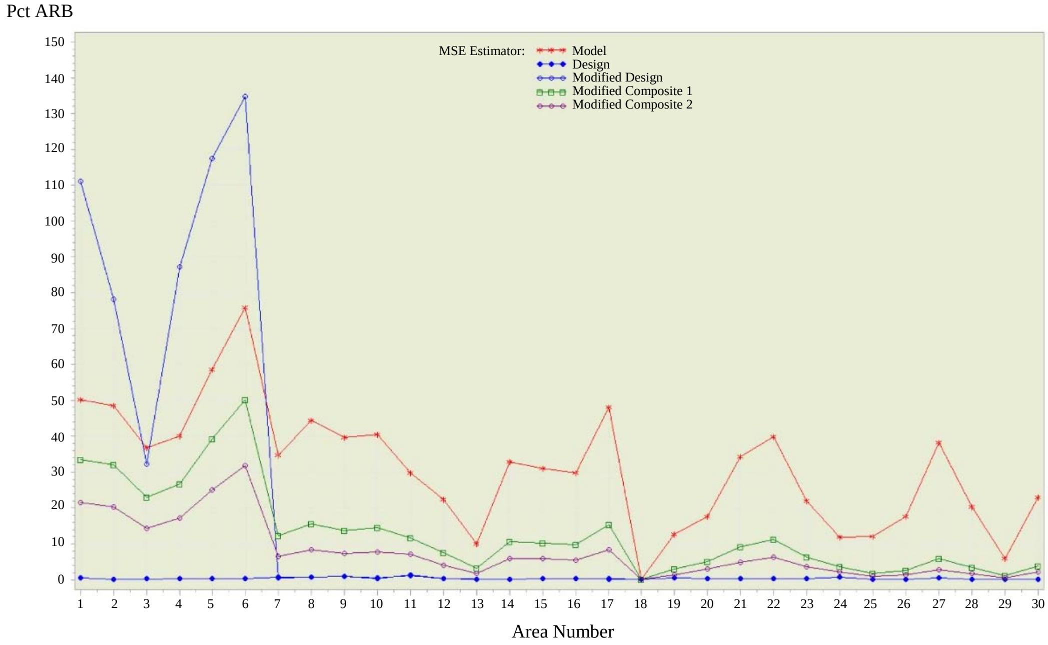


Figure 5.2 Plot of percent ARB of the MSE estimators: area level model.

Table 5.1
Mean % design ARB and mean % design RRMSE of MSE estimators: area level model

MSE Estimator	Mean % design ARB		Mean % design RRMSE	
	Areas 1 to 6	Areas 7 to 30	Areas 1 to 6	Areas 7 to 30
Design	0.33	0.39	246.71	33.62
Modified Design	93.49	0.38	221.86	33.58
Model	51.66	25.76	54.98	26.61
Modified Composite 1	34.08	7.60	96.98	24.70
Modified Composite 2	32.00	4.13	146.31	28.20

As expected, Figure 5.2 shows that the design unbiased estimator has zero ARB (except for simulation errors) across all areas. On the other hand, the modified design unbiased MSE estimator surprisingly exhibits a large ARB for the first six areas, with mean value of 93.49% but negligible for the remaining areas (0.38%). Model MSE estimator also exhibits large ARB for the first six areas with mean ARB of 51.66% that decreases to 25.76% for the remaining areas. On the other hand, the mean ARB for composite 1 MSE estimator is reduced to 34.08% for group 1 and small for group 2 (7.60%). The modified composite 2 MSE estimator that attaches more weight to the design unbiased MSE estimator reduces the mean ARB to 32.00% for group 1 and to 4.13% for group 2.

Figure 5.3 gives a plot of RRMSE of the MSE estimators across all the thirty areas and Table 5.1 reports the mean % RRMSE values for areas in group 1 and group 2. As expected, the design unbiased MSE estimator exhibits very large RRMSE for group 1 with mean value of 246.71%. The modified design unbiased MSE estimator is equally unstable for group 1 (mean RRMSE of 221.86%) in addition to exhibiting large ARB. Model MSE estimator exhibits the smallest RRMSE as expected with mean value of 54.98% for group 1 compared to 96.98% for composite 1 MSE estimator and 146.31% for modified composite 2 MSE estimator. On the other hand, for the areas in group 2 with smaller sampling variances, the mean RRMSE of the three MSE estimators is roughly the same: 24.70% for composite 1, 26.61% for model and 28.20% for modified composite 2 MSE estimators. The mean RRMSE for the design unbiased and modified design unbiased MSE estimators is only slightly larger for group 2 with values of 33.62% and 33.58% respectively.

Finally, we turn to confidence interval coverage rates for a nominal value of 95%. Normal theory coverage rates for the model MSE estimator are computed as

$$\mathrm{CR}\left[\mathrm{mse}\left(\hat{\theta}_i^{\mathrm{EB}}\right)\right] = R^{-1}\sum_{r=1}^{R} I\left[\hat{\theta}_i^{\mathrm{EB}(r)} - 1.96\left(\mathrm{mse}_i^{\mathrm{EB}(r)}\right)^{1/2} \le \theta_i \le \hat{\theta}_i^{\mathrm{EB}(r)} + 1.96\left(\mathrm{mse}_i^{\mathrm{EB}(r)}\right)^{1/2}\right] \quad (5.4)$$

where $I[\bullet]$ is an indicator function with value 1 if θ_i is in the calculated interval and 0 otherwise. Coverage rates for the other MSE estimators are similarly defined. Figure 5.4 is a plot of the percent coverage rates for the MSE estimators. The curve associated with the design-unbiased MSE estimator is not included in the plot because it is not possible to calculate the confidence interval coverage rate due to negative MSE estimates for some simulation runs. Discarding these simulation runs and calculating the intervals from the remaining runs can distort the coverage rate.

The plot shows serious undercoverage for areas in group 1 with large sampling variance. In particular, the mean coverage rate for model, modified composite 1 and modified composite 2 are 68.53%, 78.43%

and 72.87% respectively, whereas the modified design MSE estimator show some improvement: 85.82%. On the other hand, for the areas in group 2 with smaller sampling variances, the mean coverage rate increases to 91.73%, 91.74%, 90.89% and 89.85% for the model, modified composite 1, modified composite 2 and the modified design MSE estimators, respectively. Figure 5.4 suggests that the coverage rates for the model and modified composite MSE estimators are comparable across all areas with the areas in group 1 exhibiting serious undercoverage because of small sample sizes or large sampling variances in those areas.

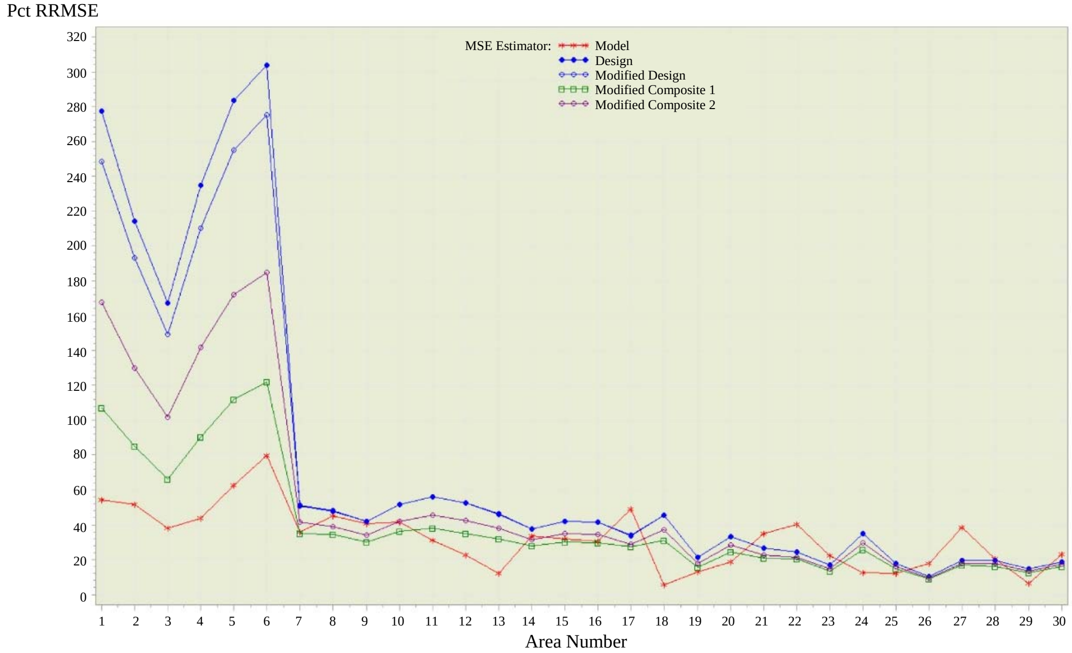


Figure 5.3 Plot of percent RRMSE of the MSE estimators: area level model.

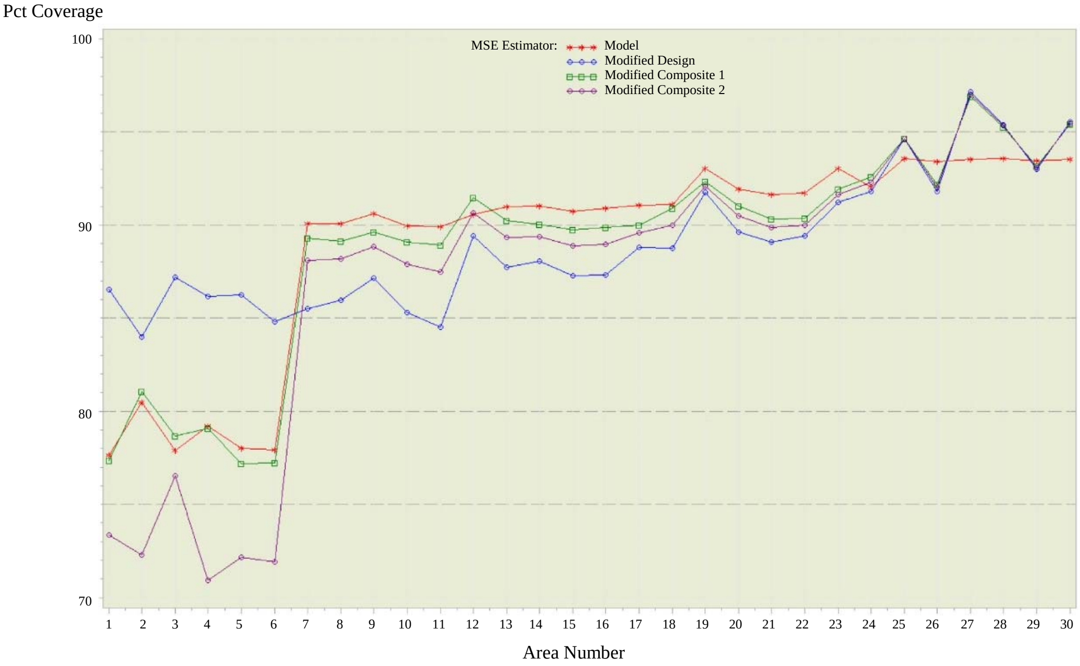


Figure 5.4 Plot of the percent coverage rates for the MSE estimators: area level model.

5.2 Unit-level model

In this section, we report some results of a limited simulation study on the design performance of four MSE estimators under a simple unit-level mean model given by

$$y_{ij} = \beta + v_i + e_{ij}, \quad j = 1, \ldots, N_i; \quad i = 1, \ldots, m \tag{5.5}$$

where the area random effects $v_i \overset{\text{iid}}{\sim} N(0, \sigma_v^2)$ are independent of the unit errors $e_{ij} \overset{\text{iid}}{\sim} N(0, \sigma_e^2)$. The MSE estimators studied include the model MSE estimator $\text{mse}\left(\hat{\bar{Y}}_i^{\text{EB}}\right)$, of the EB estimator $\hat{\bar{Y}}_i^{\text{EB}}$ (Rao and Molina, 2015, Section 7.2.3), the plug-in design-based MSE estimator $\text{mse}_d^*\left(\hat{\bar{Y}}_i^{\text{EB}}\right)$ obtained from (4.9) by replacing the model parameters β, σ_v^2 and σ_e^2 by their REML estimators, the composite MSE estimator given by (4.10), and a "conditional" MSE estimator, $\text{mse}_{\text{CH}}\left(\hat{\bar{Y}}_i^{\text{EB}}\right)$, proposed by Chambers, Chandra and Tzavidis (2001, Section 2.2.2).

For the design-based simulation, we use $m = 30$ small areas and first generate the area population sizes N_i, from a Uniform distribution $U[443, 542]$ and hold them fixed over simulation runs, following Chambers et al. (2011). We generate two fixed finite populations $\{y_{ij}, j = 1, \ldots, N_i; i = 1, \ldots, m\}$ from the mean model (5.5) for specified mean parameter $\beta = 500$ and variance parameters $\sigma_v^2 = 10.40$, $\sigma_e^2 = 94.09$ for the first finite population (denoted Population A) and $\beta = 500$, $\sigma_v^2 = 40.32$, $\sigma_e^2 = 94.09$ for the second finite population (denoted Population B). Note that the variance ratio $\delta = \sigma_v^2 / \sigma_e^2$ is equal to 0.11 for Population A and is smaller than the value 0.43 for Population B. We then draw stratified simple random samples $\{y_{ij}, j = 1, \ldots, n_i; \; i = 1, \ldots, 30\}$ without replacement, from each finite population, treating each area as a stratum, where the area sample sizes are chosen to be equal: either $n_i = 5$ or $n_i = 20$. In all, we draw $S = 10{,}000$ stratified simple random samples and compute the MSE estimates from each sample. Independently, we also draw $R = 30{,}000$ stratified random samples and compute the EB estimates from each sample. The MSE of the EB estimator for each area is approximated along the lines of (5.1) using the 30,000 simulation runs. Using a large number of simulation runs, $R = 30{,}000$, the true MSE of the EB estimator is accurately approximated by the empirical MSE. On the other hand, a smaller number of simulation runs, such as $S = 10{,}000$, is used for studying the performance of the four MSE estimators to reduce computations. This two-step simulation setup is often used for the unit level model (see e.g., González-Manteiga, Lombardia, Molina, Morales and Santamaria, 2008). Typically, calculating the MSE is much faster than calculating the RB and RRMSE of several MSE estimators, particularly bootstrap MSE estimators.

Using the simulated MSE estimates and the simulated MSE of the EB, we compute the relative bias (RB), the absolute relative bias (ARB) and the relative root mean square error (RRMSE) of the MSE estimators along the lines of (5.2) and (5.3). In the case of Population A and area sample size 5, the plug-in design-based MSE estimator leads to underestimation across all areas, with RB ranging from -87.0% to -18.1%. This underestimation is due to ignoring the variability in the parameter estimates. On the other hand, the model MSE estimator generally overestimates the design MSE with RB ranging from -66.4% to 150.1%. As a result, the composite MSE estimator reduces the underestimation caused by the plug-in design-based MSE estimator: RB ranging from -55.0% to 115.4%. The conditional MSE estimator overestimates the design MSE consistently with RB ranging from 31.7% to 316.1%. Performance of the MSE estimators in terms of RB improves as the ratio δ increases to 0.43 or the area sample size increases to 20.

Table 5.2 reports the median and mean ARB values for the two populations and the two sample sizes. It shows that the composite MSE estimator performs better than the other MSE estimators for Population A and area sample size 5, with median and mean ARB equal to 53%. On the other hand, the conditional MSE estimator exhibits large median ARB equal to 208% and mean ARB equal to 191%. Median and mean ARB values for all the MSE estimators decrease as the ratio δ increases to 0.43 or the area sample size increases to 20.

Table 5.2
Median and mean % design ARB of MSE estimators: unit level model

MSE Estimator	**Population A**				**Population B**			
	$n_i = 5$		$n_i = 20$		$n_i = 5$		$n_i = 20$	
	Median	**Mean**	**Median**	**Mean**	**Median**	**Mean**	**Median**	**Mean**
Design	60.7	54.4	11.2	11.1	8.9	8.9	1.8	2.0
Conditional	207.9	190.7	23.2	19.9	9.4	8.3	0.7	1.0
Model	77.4	81.7	44.0	38.8	29.6	28.4	6.8	8.6
Composite	52.9	53.3	13.1	14.0	7.1	8.8	1.3	1.8

Table 5.3 reports the median and mean % design RRMSE values for the two populations and the two sample sizes. It shows that the model MSE estimator and the composite MSE estimator perform better than the other MSE estimators, especially for Population A and area sample size 5. In the latter case, the plug-in design-based MSE estimator and the conditional MSE estimator exhibit large median and mean RRMSE values: approximately 400% versus 110% for the model MSE estimator and the composite MSE estimator. Performance of all the MSE estimators improves in terms of RRMSE as the ratio δ increases or the area sample size increases. In the case of population B and area sample size 20, model MSE estimator exhibits the smallest median and mean RRMSE: approximately 10% versus 30% for the other MSE estimators.

Table 5.3
Median and mean % design RRMSE of MSE estimators: unit level model

MSE Estimator	**Population A**				**Population B**			
	$n_i = 5$		$n_i = 20$		$n_i = 5$		$n_i = 20$	
	Median	**Mean**	**Median**	**Mean**	**Median**	**Mean**	**Median**	**Mean**
Design	414.5	382.0	62.1	60.3	57.6	57.6	29.3	29.0
Conditional	416.5	384.5	64.1	62.2	63.9	64.3	28.4	28.1
Model	107.8	108.5	45.4	41.6	31.6	31.7	8.9	10.8
Composite	113.7	112.9	37.8	38.1	40.7	41.5	26.6	26.4

6 Conclusions

In this paper we studied the properties of alternative MSE estimators in tracking the design MSE of EB estimators of small area means. We examined both area level and unit level models.

In the area level model, we proposed two composite MSE estimators by taking a weighted average of a design unbiased MSE estimator and a model based MSE estimator. Modifications to ensure positive MSE estimators were also given. Performance of the alternative MSE estimators was studied through simulations in terms of absolute relative bias, relative root mean square error and coverage rate of confidence intervals. Our results for the area level model suggest that the design unbiased MSE estimator is not usable in practice when the area sample size is very small because of a large probability of getting a negative value. On the other hand, this probability for the composite 1 MSE estimator (with the same weights as the EB estimator), is either zero or essentially negligible. Our simulations for the area level model for areas with very small sample sizes suggest that the composite 1 MSE estimator leads to smaller ARB relative to the model MSE estimator at the expense of an increase in RRMSE. For areas with larger sample sizes, the ARB of the model MSE estimator persists unlike the ARB of the composite 1 MSE estimator. In terms of coverage rates, the model MSE estimator and the composite 1 MSE estimator are comparable across all areas, but both can lead to serious undercoverage for areas with very small sample sizes. Overall, the composite 1 MSE estimator provides a good compromise in estimating the design MSE.

In the simulation study of the unit level model, our results suggest that the composite MSE estimator generally offers a good compromise between the ARB and RRMSE. However, the plug-in design MSE estimator used in the composite estimator needs modification to take account of the variability in the estimators of model parameters to avoid or reduce the underestimation of design MSE of the EB estimator.

Acknowledgements

The authors thank the assistant and associate editors and the two referees for their comments and suggestions that led to improvements in our examination of the various MSE estimators.

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.

Beaumont, J.-F., and Bocci, C. (2016). Small area estimation in the Labour Force Survey. Unpublished manuscript.

Chambers, R., Chandra, H. and Tzavidis, N. (2011). On bias-robust mean squared error estimation for pseudo-linear small area estimators. *Survey Methodology*, 37, 2, 153-170. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2011002/article/11604-eng.pdf.

Datta, G., 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.

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

González-Manteiga, W., Lombardia, M.J., Molina, I., Morales, D. and Santamaria, L. (2008). Bootstrap mean squared error of a small-area EBLUP. *Journal of Statistical Computation and Simulation*, 78, 443-462.

Molina, I., and Kominiak, E.S. (2017). Estimation of proportions in small areas: Application to the labour force using the Swiss Census Structural Survey. Unpublished technical report.

Pfeffermann, D., and Gilboa, A. (2017). Estimation of randomization MSE in small area estimation. Paper presented at the 2017 SAE conference, Paris.

Rao, J.N.K., and Molina, I. (2015). *Small Area Estimation, Second Edition.* Hoboken, New Jersey: Wiley.

Rivest, L.-P., and Belmonte, E. (2000). A conditional mean squared error of small area estimators. *Survey Methodology*, 26, 1, 67-78. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2000001/article/5179-eng.pdf.

Rubin-Bleuer, S., and Schiopu-Kratina, I. (2005). On the two-phase framework for joint model and design-based inference. *The Annals of Statistics*, 33, 6, 2789-2810.

Small area estimation for unemployment using latent Markov models

Gaia Bertarelli, M. Giovanna Ranalli, Francesco Bartolucci, Michele D'Alò and Fabrizio Solari[1]

Abstract

In Italy, the Labor Force Survey (LFS) is conducted quarterly by the National Statistical Institute (ISTAT) to produce estimates of the labor force status of the population at different geographical levels. In particular, ISTAT provides LFS estimates of employed and unemployed counts for local Labor Market Areas (LMAs). LMAs are 611 sub-regional clusters of municipalities and are unplanned domains for which direct estimates have overly large sampling errors. This implies the need of Small Area Estimation (SAE) methods. In this paper we develop a new area level SAE method that uses a Latent Markov Model (LMM) as linking model. In LMMs, the characteristic of interest, and its evolution in time, is represented by a latent process that follows a Markov chain, usually of first order. Therefore, areas are allowed to change their latent state across time. The proposed model is applied to quarterly data from the LFS for the period 2004 to 2014 and fitted within a hierarchical Bayesian framework using a data augmentation Gibbs sampler. Estimates are compared with those obtained by the classical Fay-Herriot model, by a time-series area level SAE model, and on the basis of data coming from the 2011 Population Census.

Key Words: Area level model; Hierarchical Bayes; Time-series data; Labor Force Survey; Augmented data.

1 Introduction

In Italy, the Labor Force Survey (LFS) is conducted quarterly by ISTAT, the National Statistical Institute, to produce estimates of the labor force status of the population at a national, regional (NUTS2), and provincial (LAU1) level, with monthly, quarterly, and yearly frequency, respectively. Since 1996, ISTAT also disseminates yearly LFS estimates of employed and unemployed counts for local Labor Market Areas (LMAs). LMAs are sub-regional geographical areas where the bulk of the labor force lives and works, and where establishments can find the largest amount of the labor force necessary to occupy the offered jobs. These are 611 distinct and functional areas defined as clusters of municipalities through an allocation process based on commuting patterns collected by the 2011 Population Census (Istat, 2014). Unlike NUTS2 and LAU1 areas, LMAs are unplanned domains that cut across sampling strata and LAU1 areas. In addition, direct estimators have overly large sampling errors particularly for areas with small sample sizes. This makes it necessary to borrow strength from data on auxiliary variables from other areas through appropriate models, leading to indirect or model-based estimates.

Small Area Estimation (SAE) methods are used in inference for finite populations to obtain estimates of parameters of interest when domain sample sizes are too small to provide adequate precision for direct domain estimators. Statistical models for SAE can be formulated at the individual or area (i.e., aggregate) levels. In this paper we focus on the latter. The Fay-Herriot model (Fay and Herriot, 1979, FH) is the basic area level SAE model: it uses cross-sectional information for predicting small area parameters of interest by combining direct estimates and population level auxiliary information with a linear mixed model. When

[1] Gaia Bertarelli, Dept. of Economics and Management, University of Pisa, Italy; M. Giovanna Ranalli, Dept. of Political Science, University of Perugia, Italy. E-mail: giovanna.ranalli@unipg.it; Francesco Bartolucci, Dept. of Economics, University of Perugia, Italy; Michele D'Alò and Fabrizio Solari, Italian National Statistical Institute.

longitudinal data are also available, it is possible to borrow strength over time. Among others, Rao and Yu (1994) propose a model involving autocorrelated random effects and use both time-series and cross-sectional data, while Marhuenda, Molina and Morales (2013) develop a spatio-temporal FH model using an autoregressive model in space together with a first-order autoregressive covariance structure in time.

Several papers deal with SAE using time-series models and the Kalman filter after expressing them in a state-space form. Pfeffermann and Burck (1990) introduce state-space models to estimate the Canadian unemployment rates and Pfeffermann and Rubin-Bleuer (1993) use this approach to model the correlation between the trends of domain series in a multivariate structural time-series model. Pfeffermann and Tiller (2006) add monthly benchmark constraints to the time-series state-space model, while Harvey and Chung (2000) consider a bivariate state-space model to obtain more stable and precise estimates of change in unemployment. Krieg and Van der Brakel (2012) model domain series in a multivariate time-series model and apply the cointegration idea to construct more parsimonious common trend models. Level break estimation within the structural time-series framework is illustrated in Van den Brakel and Krieg (2015). More recently, Van der Brakel and Krieg (2016) and Boonstra and Van den Brakel (2016) apply these models to data from the Dutch LFS.

Proposals for area level time-series data have also been developed following a Hierarchical Bayesian (HB) approach. In particular, Ghosh, Nangia and Kim (1996) apply a fully HB analysis using a time-series model to the estimation of median income of four-person families. Datta, Lahiri, Maiti and Lu (1999) apply this approach to a longer time-series from the U.S. Current Population Survey and use a random walk model for the area random effects. You, Rao and Gambino (2003) apply the same model to unemployment rate estimation for the Canadian LFS. Recently, Boonstra (2014) uses a time-series HB multilevel model to estimate unemployment at the municipality level using data from the Dutch LFS. In particular, estimates are obtained for each quarter and include random municipality effects and random municipality by quarter effects.

In this work we develop a new area level SAE method based on Latent Markov Models (LMMs, see Bartolucci, Farcomeni and Pennoni, 2013, for a thorough description) to estimate unemployment incidences in LMAs using quarterly data from 2004 to 2014 within an HB framework. Area level SAE models consist of two parts, a sampling model formalizing the assumptions on direct estimators and their relationship with underlying area parameters, and a linking model that relates these parameters to area specific auxiliary information. In this work, an LMM is used as linking model and the sampling model is introduced as the highest level of the hierarchy. The resulting model is fitted within a Bayesian framework using a Gibbs sampler with augmented data (corresponding to the latent variables) that allows for a more efficient sampling of the model parameters (Tanner and Wong, 1987).

LMMs, introduced by Wiggins (1973), allow for the analysis of longitudinal data when the response variables measure common characteristics of interest that are not directly observable. The basic LMM formulation is similar to that of hidden Markov models for time-series data (MacDonald and Zucchini, 1997). In these models, the characteristics of interest and their evolution in time are represented by a latent process that follows a Markov chain, typically of first order, so that single areas are allowed to move between latent states across time. LMMs may be seen as an extension of Markov chain models to control for measurement errors. Moreover, LMMs can be seen as an extension of latent class models (Lazarsfeld, Henry and Anderson, 1968) to longitudinal data. Latent class models have been considered in a SAE framework in Fabrizi, Montanari and Ranalli (2016), where a latent class unit level model for predicting disability small area counts from survey data is introduced for cross sectional data.

The remainder of this paper is organized as follows. Section 2 provides a more detailed description of the available LFS data, while Section 3 introduces notation and reviews some relevant time-series area level SAE methods available in the literature. In Section 4, the model and the procedure for its estimation are presented in detail. Section 5 is devoted to the discussion of the results of the application to the LFS data. Conclusions and possible future developments are outlined in Section 6.

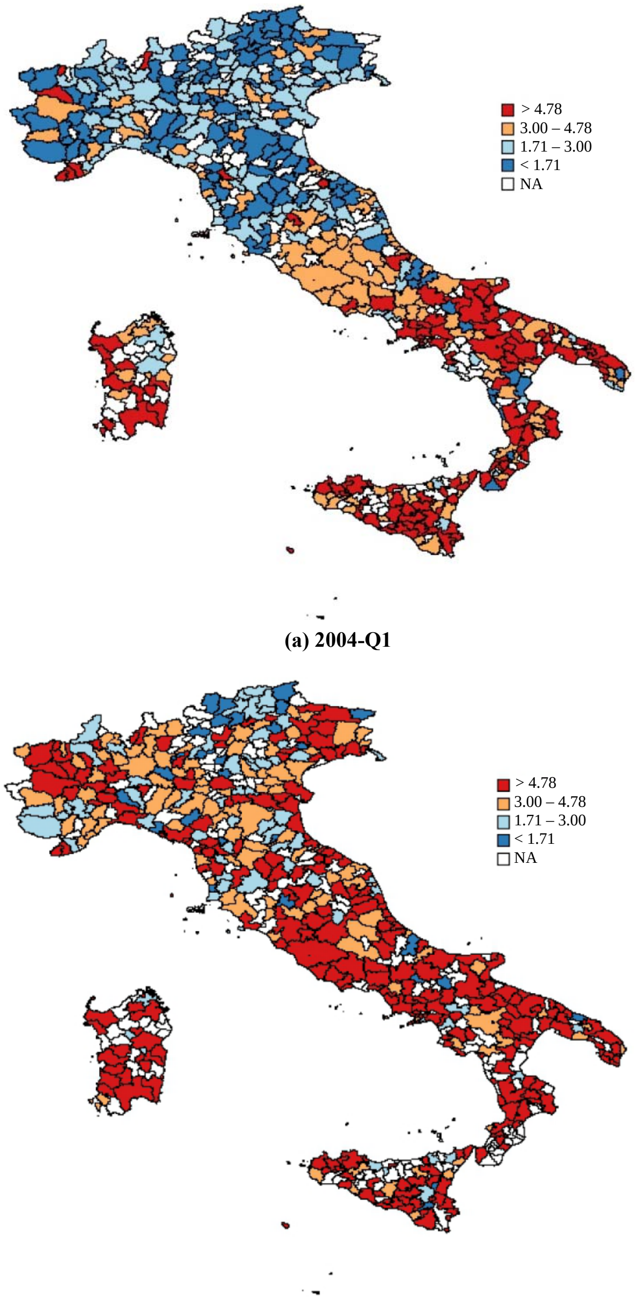
2 Data and preliminary analysis

As already mentioned, LMAs are unplanned domains for the LFS. In fact, the sampling design is as follows. Within a given LAU1, municipalities are classified as Self-Representing Areas (larger municipalities) and Non-Self-Representing Areas (smaller municipalities). In Self-Representing Areas, a stratified cluster sampling design is applied: each municipality is a single stratum and households are selected by means of systematic sampling. In Non-Self-Representing Areas, the sample is based on a stratified two stage sampling design: municipalities are Primary Sampling Units, while households are Secondary Sampling Units. Primary Sampling Units are divided into strata of the same dimension in terms of population size. One Primary Sampling Unit is drawn from each stratum without replacement and with probability proportional to the Primary Sampling Unit population size. Secondary Sampling Units are selected by means of systematic sampling in each Primary Sampling Unit. All members of each sample household, both in Self-Representing Areas and in Non-Self-Representing Areas are interviewed. In each quarter, about 70,000 households and 1,350 municipalities are included in the sample. Note that some LMAs (usually the smallest ones) have a very small sample size. Furthermore, usually about one third of the LMAs is not included in the sample at all (i.e., they have a zero sample size).

The LFS follows a rotating panel sampling design, according to a 2-(2)-2 scheme: households are interviewed in two consecutive quarters and, after a two-quarter break, they are interviewed for two additional consecutive quarters. Although the LFS panel design induces correlation among quarterly estimates, due to partial overlap of the sample units, we do not account for it in our model specification in the application illustrated in Section 5. In any case, we expect that this does not affect the comparison among different methods.

In this work we model quarterly unemployment incidences for 611 LMAs for the period 2004-Q1 to 2014-Q4 (44 quarters). Figure 2.1 shows the map of direct estimates in the first and in the last time occasion of the observation time span. Figure 2.2, on the other hand, shows all the direct estimates for each small area in two NUTS2 areas: Lombardy (left panel) is a rich region in the North of Italy, while Sicily (right panel) is the southern Island and is much less wealthy. We observe, in general, that direct estimates are extremely variable and that unemployment has decreased over the first three years, and then started to increase considerably.

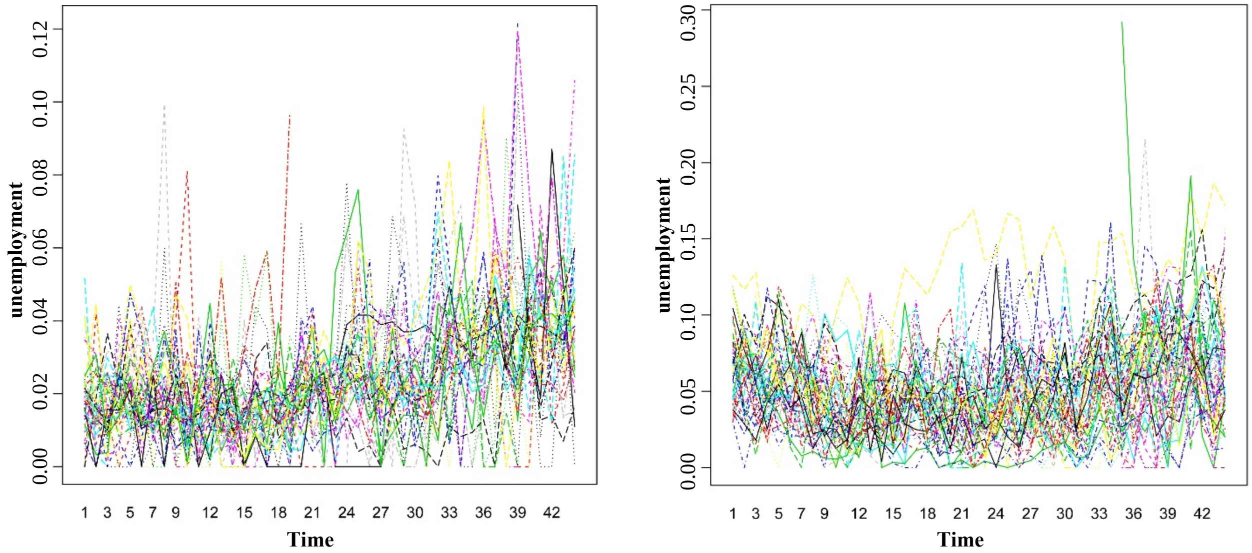
Direct estimates in unplanned domains are characterized by a high Coefficient of Variation (CV), which is used as a measure of uncertainty associated with the estimates. In addition, 6,762 out of 26,884 direct estimates cannot be computed because the sample dimension is zero. Usually, in Official Statistics, an estimate for a Labor Force parameter with a CV greater than 33.3% is considered too unreliable and is not recommended for release. Estimates with a CV between 16.6% and 33.3% must be released with caveats because their sampling variability is quite high, while estimates with a CV smaller than 16.6% are of sufficient accuracy and have no release restrictions; see Statistics Canada (2016, page 35). In our data, the vast majority of direct estimates has a very large CV and cannot be considered reliable, as it is shown in Figure 2.3.



(a) 2004-Q1

(b) 2014-Q4

Figure 2.1 Direct estimates of unemployment incidences (%) for the first and the last time occasion: first quarter of 2004 (a) and the last quarter of 2014 (b).



(a) Lombardy **(b) Sicily**

Figure 2.2 Quarterly direct estimates of unemployment incidences in two NUTS2 Regions: Lombardy (a) and Sicily (b), from 2004-Q1 to 2014-Q4.

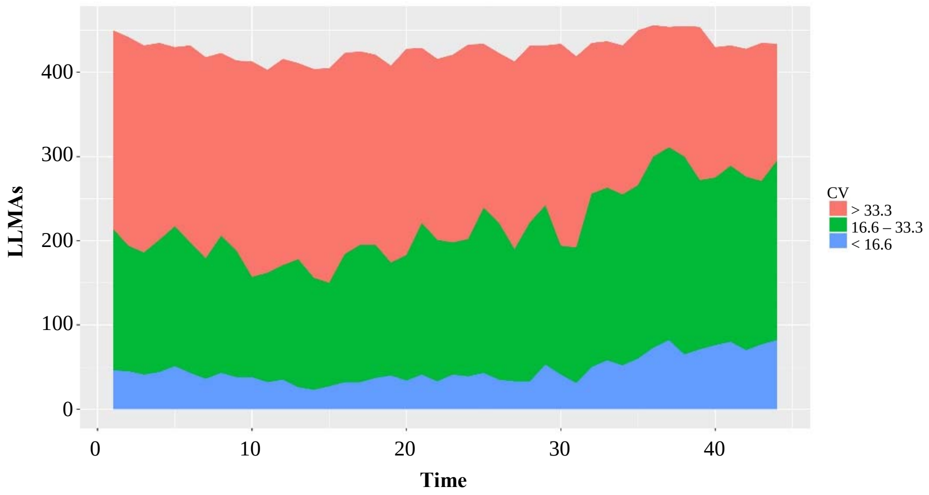


Figure 2.3 For each quarter, distribution of the sampled small areas according to classes of values of the CV of the direct estimates.

The basic idea of SAE is to introduce a statistical model to exploit the relationship between the variable of interest and some covariates for which population information is available. Auxiliary variables available for these data are the population rates in $sex \times 7\, age$ classes (15-19, 20-24, 25-29, 30-39, 40-49, 50-59, 60-74). Since LFS estimates are not seasonally adjusted, we take seasonality into account using year and quarter effects through 10 and 3 dummy variables, respectively.

The CVs of direct estimates are estimates themselves and their precision is a function of the sample size. Therefore, they are subject to a relevant sampling error that can affect small area modeling in different ways (Rao and Yu, 1994) and smoothing estimated Mean Squared Errors (MSEs) is necessary (see Rao, 2003, Chapter 5). In this work, we propose to use a regression model with a logarithmic transformation of the CV and of the MSE (see Wolter, 2007, Chapter 7). In particular, our approach is based on two steps: the first step consists in modeling the CV and then computing the smoothed MSE from this model. At the second step, we model the MSE directly for those small areas for which we do not have a valid CV (i.e., for those LMAs with a zero estimate).

Let $\hat{\theta}_{it}$ be the direct survey estimate for small area $i = 1, \ldots, m,$ with $m = 611,$ at time $t = 1, \ldots, T,$ with $T = 44.$ Let CV_{it} denote the corresponding estimate of the CV. Note that Italy is divided into four geographic areas, namely broad-areas (e.g., North-West, North-East, Center, South and Islands), and that each LMA belongs to only one of these broad-areas. In order to smooth estimates of MSEs, we have the following auxiliary information:

- M_{it} is the population size at time t of the broad-area to which LMA i belongs;
- N_{it} is the population size of LMA i at time t;
- r_{it} is a 14-dimensional column-vector that contains population rates in $sex \times 7age$ classes, for LMA i at time t.

At the first step of the proposed procedure, we fit the following regression model for each broad-area:

$$\log\left(\mathrm{CV}_{it}\right) = \beta_0 + \log\left(\hat{\theta}_{it}\right)\beta_1 + \log\left(\frac{N_{it}}{M_{it}}\right)\beta_2 + \log\left(\mathbf{r}_{it}\right)'\boldsymbol{\beta}_3 + \log\left(\mathbf{1}_{14} - \mathbf{r}_{it}\right)'\boldsymbol{\beta}_4, \qquad (2.1)$$

where $\mathbf{1}_{14}$ is a 14-dimensional column vector of ones. The use of the log-transformation and the choice of the covariates has been assessed using standard model selection techniques, such as AIC and adjusted R^2. Using predictions denoted by $\widehat{\mathrm{CV}}_{it}$ from this model, smoothed MSEs are obtained as

$$\widehat{\mathrm{MSE}}_{it} = \widehat{\mathrm{CV}}_{it} \times \hat{\theta}_{it}.$$

In the second step of the proposed procedure, for all $\hat{\theta}_{it} = 0,$ CVs cannot be computed while MSEs are available since direct estimates are based on calibrated weights and MSE estimates are based on the residuals of a generalized regression that accounts for the auxiliary variables used in the calibration constraints. Then, MSEs are modeled directly and separately for each broad-area using the following model:

$$\log\left(\mathrm{MSE}_{it}\right) = \beta_0 + \log\left(\frac{N_{it}}{M_{it}}\right)\beta_1 + \log\left(\mathbf{r}_{it}\right)'\boldsymbol{\beta}_2 + \log\left(\mathbf{1}_{14} - \mathbf{r}_{it}\right)'\boldsymbol{\beta}_3.$$

Smoothed MSEs are obtained as predictions from this model. Note that, we have resorted to this two-step procedure because the former model, the one for CVs, fitted better than the latter for MSEs in our application. Figure 2.4 reports the final output of this two-step procedure and displays the original and the smoothed MSEs versus unemployment incidence for all sampled areas.

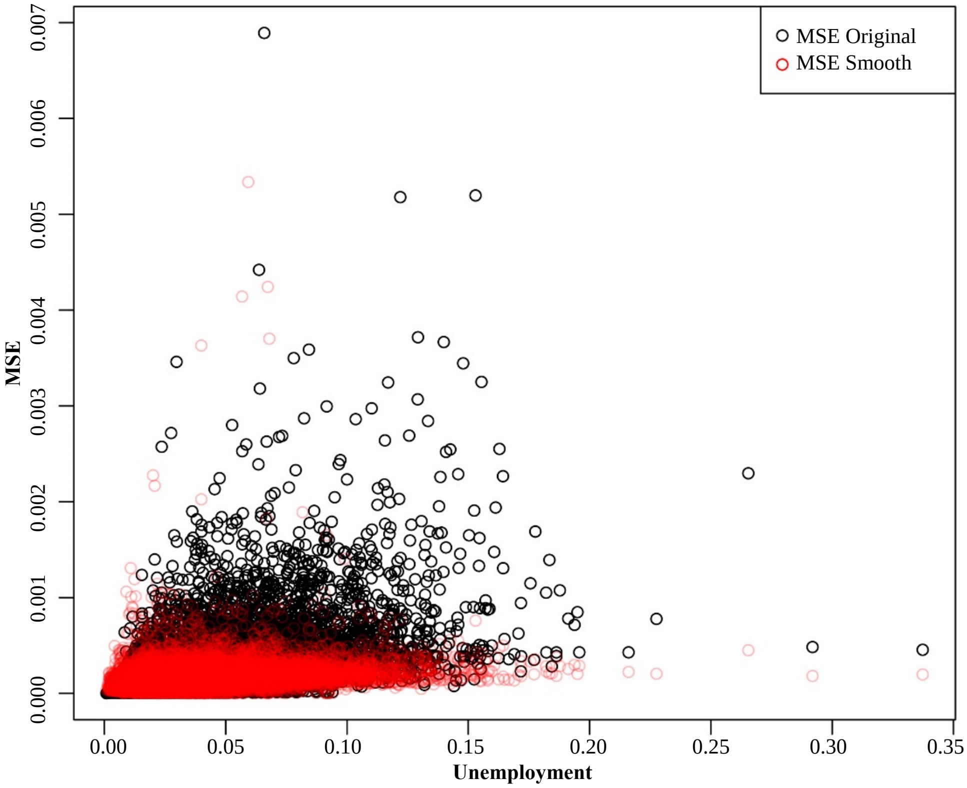


Figure 2.4 Original (black) and smoothed (red) MSEs vs unemployment incidence for all sampled areas.

3 Time series area level SAE models

Rao and Yu (1994) propose an area level model involving autocorrelated random effects and sampling errors using both time-series and cross sectional data. It consists of a sampling model

$$\hat{\theta}_{it} = \theta_{it} + e_{it}, \quad i = 1, \ldots, m, \; t = 1, \ldots, T,$$

and an area-linking model

$$\theta_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + v_i + u_{it}, \quad i = 1, \ldots, m, \; t = 1, \ldots, T,$$

where θ_{it} is the true value corresponding to the estimate $\hat{\theta}_{it}$ for the small area mean, $\mathbf{x}_{it}$ is a $p-$ dimensional column vector of fixed covariates, and e_{it} are normal sampling errors. Given the true value θ_{it}, each vector $\mathbf{e}_i = (e_{i1}, \ldots, e_{iT})'$ has multivariate normal distribution with zero mean and with known variance-covariance matrix $\boldsymbol{\Psi}_i$. Moreover, $v_i \sim N(0, \sigma_v^2)$ is the area effect and $u_{it} = \rho u_{i,t-1} + \epsilon_{it}$, with $|\rho| < 1$ and $\epsilon_{it} \sim N(0, \sigma_\epsilon^2)$ is the area-by-time effect. In this model, $\mathbf{e}_i, v_i$, and ϵ_{it} are assumed independent of each other. In our application $\boldsymbol{\Psi}_i$ is diagonal, with elements ψ_{it}, for $t = 1, \ldots, T$.

In the previous formulation, the area-linking model is basically a linear model with mixed coefficients. You et al. (2003, YRG) translate this model into an HB framework as follows. Let $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{iT})'$ and $\hat{\boldsymbol{\theta}}_i = \left(\hat{\theta}_{i1}, \ldots, \hat{\theta}_{iT}\right)'$, then

$$\begin{aligned}\hat{\boldsymbol{\theta}}_i \,|\, \boldsymbol{\theta}_i &\sim N_T\left(\boldsymbol{\theta}_i,\, \boldsymbol{\Psi}_i\right),\\ \theta_{it} \,|\, \boldsymbol{\beta},\, u_{it},\, \sigma_v^2 &\sim N\left(\mathbf{x}_{it}'\boldsymbol{\beta} + u_{it},\, \sigma_v^2\right),\\ u_{it} \,|\, u_{i,t-1},\, \sigma_\epsilon^2 &\sim N\left(\rho u_{i,t-1},\, \sigma_\epsilon^2\right),\end{aligned} \tag{3.1}$$

where $\boldsymbol{\beta}$, σ_v^2, and σ_ϵ^2 are mutually independent. The model is fully specified once priors are chosen for $\boldsymbol{\beta}$, σ_v^2, and σ_ϵ^2, namely as $f(\boldsymbol{\beta}) \propto 1$, $\sigma_v^2 \sim \text{IG}(a_1, b_1)$, and $\sigma_\epsilon^2 \sim \text{IG}(a_2, b_2)$, where a_1, a_2, b_1 and b_2 are known positive hyperparameters and, usually, set to be small and to reflect a vague knowledge about σ_v^2 and σ_ϵ^2.

Datta et al. (1999) follow this approach, but introduce a richer structure for the fixed part of the linking model by assuming

$$\theta_{it} = \mathbf{x}_{it}'\boldsymbol{\beta}_i + v_i + u_{it}, \tag{3.2}$$

where v_i and $\boldsymbol{\beta}_i$ are area-specific intercepts and regression coefficients, respectively, and u_{it} is an area-specific error term that follows the random-walk model

$$u_{it} \,|\, u_{i,t-1},\, \sigma_\epsilon^2 \sim N\left(u_{i,t-1},\, \sigma_\epsilon^2\right).$$

The column vector of auxiliary variables $\mathbf{x}_{it}$ may also include dummy variables for year and/or seasonality adjustments. Note that area-specific regression coefficients considerably increase the estimation complexity and the computational burden. For this reason, the hyperparameters are assumed to be m independent realizations from a common probability distribution specified by $v_i \sim N(0, \sigma_v^2)$ and $\boldsymbol{\beta}_i \sim N\left(\boldsymbol{\beta}, \mathbf{W}_\beta^{-1}\right)$, which, in turn, depend on appropriate parameters. See Datta et al. (1999) for further details.

4 The proposed model

In this section, the proposed SAE model based on LMMs is illustrated. It can be considered as a compromise between the YRG model based on (3.1), which leads to possible oversmoothing, and the computationally demanding alternative proposed in Datta et al. (1999), based on (3.2). We first outline a general description on LMMs and then move to the specification of the area level model and to its estimation.

4.1 Preliminaries

In LMMs, the existence of two types of process is assumed: an unobservable finite-state first-order Markov chain U_{it} with state space $\{1, \ldots, k\}$ and an observed process, which in our case corresponds to θ_{it}, with $i = 1, \ldots, m$ and $t = 1, \ldots, T$. It is assumed that the distribution of θ_{it} depends only on U_{it}; specifically, the θ_{it} are conditionally independent given the U_{it}. In addition, the latent state to which a small area belongs at a certain time point only depends on the latent state at the previous occasion.

The state-dependent distribution, namely the distribution of θ_{it} given U_{it}, can be a continuous or discrete. Such a distribution is typically taken from the exponential family. Thus, the overall vector of parameters of LMM, denoted by $\boldsymbol{\phi}$, includes parameters of the Markov chain, denoted by $\boldsymbol{\phi}_{\text{lat}}$, and the vector of parameters $\boldsymbol{\phi}_{\text{obs}}$ of the state-dependent distribution. In fact, the model consists of two components, the measurement model and the latent model, which concern the conditional distribution of the response variables given the latent variables and the distribution of the latent variables, respectively. By jointly considering these components, the so-called manifest distribution is obtained: it is the marginal distribution of the response variables, once the latent variables have been integrated out.

The measurement model, based on parameters $\boldsymbol{\phi}_{\text{obs}}$, can be written as

$$\theta_{it} \,|\, U_{it} = u \sim p(\theta_{it} \,|\, u, \boldsymbol{\phi}_{\text{obs}}).$$

Moreover, the parameters $\boldsymbol{\phi}_{\text{lat}}$ of the Markov chain are:

- the vector of initial probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)'$ where

$$\pi_u = P(U_{i1} = u), \quad u = 1, \ldots, k;$$

- the transition probability matrix

$$\boldsymbol{\Pi} = \begin{pmatrix} \pi_{1|1} & \ldots & \pi_{1|k} \\ \vdots & \ddots & \vdots \\ \pi_{k|1} & \ldots & \pi_{k|k} \end{pmatrix},$$

where

$$\pi_{u|\bar{u}} = P(U_{it} = u \,|\, U_{i,t-1} = \bar{u}), \quad \bar{u}, u = 1, \ldots, k,$$

is the probability that area i visits state u at time t given that at time $t-1$ it was in state $\bar{u}$.

In this work we consider homogeneous LMMs, namely LMMs where, in agreement with the previous definition, the transition probability matrix is constant in time. Generalizations to non-homogeneous hidden Markov chains and time-varying transition probabilities could also be considered (Bartolucci and Farcomeni, 2009). Individual covariates could be included in the measurement or in the latent model. When the covariates are included in the measurement model (Bartolucci and Farcomeni, 2009), they affect the response variables directly and the latent process is conceived as a way to account for the unobserved heterogeneity between areas. Differently, when the covariates are in the latent model (Vermunt and Magidson, 2002; Bartolucci, Pennoni and Francis, 2007) they influence initial and transition probabilities of the latent process. In a SAE context, we will consider the former approach, so that auxiliary information can be used to improve predictions. Bayesian inference approaches to LMMs are already available in the literature (e.g., in Marin, Mengersen and Robert, 2005; Spezia, 2010). In the following section we illustrate how to incorporate an LMM into an area level SAE model.

4.2 Proposed approach to area level SAE

The proposed model is based on two levels in an HB framework: at the first level, a sampling error model is assumed, then an LMM is used as linking model. The latter is based on two equations, corresponding to the measurement model and to the latent component. In particular, we adopt the following structure:

- Sampling Model:

$$\hat{\boldsymbol{\theta}}_i \,|\, \boldsymbol{\theta}_i \sim N_T\left(\boldsymbol{\theta}_i, \boldsymbol{\Psi}_i\right), \quad i = 1, \ldots, m;$$

- Linking Model:
  - Measurement Model:

$$\theta_{it} \,|\, U_{it} = u, \mathbf{x}_{it} \sim N\left(\mathbf{x}'_{it}\boldsymbol{\beta}_u, \sigma_u^2\right) \quad i = 1, \ldots, m; \; t = 1, \ldots, T;$$

  - Latent Model, based on the initial probabilities $\pi_u, u = 1, \ldots, k,$ and on the transition probabilities $\pi_{u|\bar{u}}, t = 2, \ldots, T, \bar{u}, u = 1, \ldots, k,$ already defined.

Here $\boldsymbol{\beta}_u$ is the $p \times 1$ vector of the regression coefficients for the latent state to which area i at time t belongs, σ_u^2 is the corresponding error variance, and $\boldsymbol{\Psi}_i$ is the matrix of sampling variances, which is assumed to be known.

It must be noticed that, while in the classical area level SAE models heterogeneity is modeled using continuous (usually Normally distributed) random variables, here it is modeled with a discrete dynamic variable. As we can deduce from Figure 4.1, our data have a skewed distribution. However, the empirical distribution is not far from a Normal distribution. D'Alò, Di Consiglio, Falorsi, Ranalli and Solari (2012) show that the differences in estimates between adopting a Normal or a Binomial model are not as relevant as expected and Normal models are often used for estimation of unemployment rates (You et al., 2003; Boonstra, 2014). Finally adopting the Normal distribution has computational advantages which are clarified later in this section.

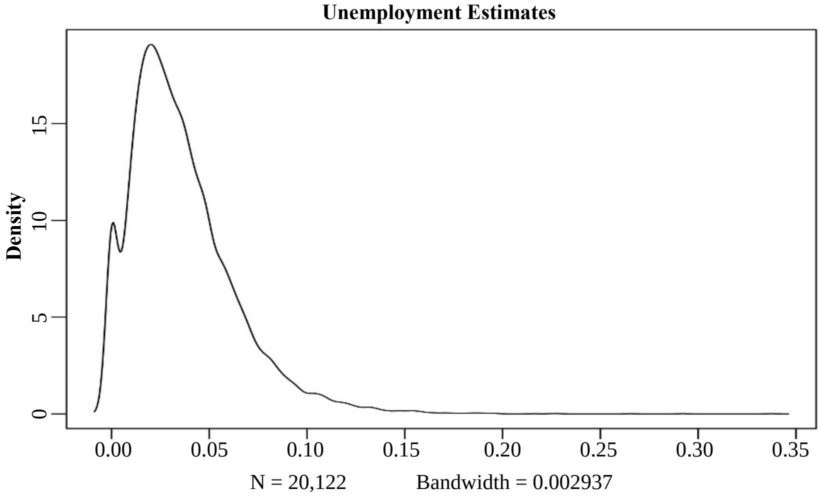


Figure 4.1 Density kernel plot of the direct estimates of unemployment incidences.

The model parameters of interest can be divided into three groups:

- the matrix of small area parameters:

$$\boldsymbol{\Theta} = \begin{pmatrix} \theta_{11} & \cdots & \theta_{1T} \\ \vdots & \ddots & \vdots \\ \theta_{m1} & \cdots & \theta_{mT} \end{pmatrix}; \tag{4.1}$$

- the vector of the measurement parameters:

$$\boldsymbol{\phi}_{\text{obs}} = \left(\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_k', \sigma_1^2, \ldots, \sigma_k^2\right)';$$

- the set of latent parameters:

$$\boldsymbol{\phi}_{\text{lat}} = \{\boldsymbol{\pi}, \boldsymbol{\Pi}\}.$$

To complete the Bayesian formulation of the proposed model, it is necessary to choose priors for the model parameters. Small area parameters do not need a specific prior because direct estimates based on observed data are available; therefore, a set of priors is chosen for the measurement and the latent parameters. Regarding $\boldsymbol{\phi}_{\text{obs}}$, diffuse normal priors are assumed for the regression coefficients. These priors are conjugate and computationally more convenient than the usually flat priors over the real line (see Rao, 2003, Chapter 10). In particular, we assume

$$\boldsymbol{\beta}_u \sim N_p\left(\boldsymbol{\eta}_0, \boldsymbol{\Sigma}_0\right), \quad u = 1, \ldots, k,$$

with $\boldsymbol{\Sigma}_0 = \sigma_u^2 \boldsymbol{\Lambda}_0^{-1}$ and $\boldsymbol{\Lambda}_0$ is a known diagonal matrix.

Variances $\sigma_u^2, u = 1, \ldots, k,$ are unknown and, therefore, it is necessary to set a prior also on these parameters. The choice of the prior distribution for the variance components is critical as in Bayesian mixed models the posterior distributions of these parameters are known to be sensitive to this specification. The inverse Gamma distribution is a popular choice, see e.g., You et al. (2003) and Datta, Lahiri, Maiti and Lu (1999) among others. Gelman (2006), Gelman, Jakulin, Pittau and Su (2008), and Polson and Scott (2012) propose to assume a half-Cauchy distribution for the variance of the random effect. Alternatively, a Uniform distribution can also be considered. Fabrizi et al. (2016) conduct an exhaustive sensitivity analysis when using a latent class model in a multivariate setting and find no significant difference among these different alternatives. For this reason, we choose the same prior distribution considered in You et al. (2003) and use an inverse Gamma distribution with shape parameter a_0 and scale parameter b_0; then $\sigma_u^2 \sim \text{IG}\left(a_0, b_0\right)$, $u = 1, \ldots, k,$ where $a_0, b_0 > 0$ are set to very small values. This choice makes it also easier to derive the full conditional distributions for the Gibbs sampler.

For $\boldsymbol{\phi}_{\text{lat}}$, a system of Dirichlet priors is set on the initial probabilities and on the transition probabilities. The Dirichlet distribution is a conjugate prior for the multinomial distribution. This means that if the prior distribution of the multinomial parameters is Dirichlet then the posterior distribution belongs to the same family. The benefit of this choice is that the posterior distribution is easy to compute and, in some sense, it is possible to quantify how much our beliefs have changed after collecting the data. Then, we assume

$$\begin{aligned}\boldsymbol{\pi} &\sim \text{Dirichlet}\left(\mathbf{1}_k\right),\\ \boldsymbol{\pi}_{\bar{u}} = \left(\pi_{1|\bar{u}}, \ldots, \pi_{k|\bar{u}}\right)' &\sim \text{Dirichlet}\left(\mathbf{1}_k\right), \quad \bar{u} = 1, \ldots, k.\end{aligned}$$

4.3 Estimation and model selection

In this work we make use of a data augmentation Markov Chain Monte Carlo (MCMC) method (Tanner and Wong, 1987; Liu, Wong and Kong, 1994; Van Dyk and Meng, 2001) based on the Gibbs sampler, in which the latent variables are treated as missing data (Marin et al., 2005; Germain, 2010). There are two main reasons for this choice. First of all, there is evidence that data augmentation has a better performance than other methods, as the marginal updating scheme (Boys and Henderson, 2003). Moreover, it simplifies the process of sampling from the posterior distribution. Details on this method and the full conditionals employed in the Gibbs sampler are given in Appendix A.1.

The choice of the number of latent states is a crucial step in applications. In the framework of LMMs, this requires a model selection procedure. From a Bayesian perspective, a fundamental goal is the computation of the marginal likelihood of the data for a given model. In this paper we use a model selection method based on the marginal likelihood and to estimate this quantity we use the method proposed by Carlin and Chib (1995), applied for each available model on the basis of the output of the MCMC algorithm. Technical details are provided in Appendix A.2.

A well-known problem occurring in Bayesian latent class and LMMs is the label switching. This implies that the component parameters are not identifiable as they are exchangeable. In a Bayesian context, if the prior distribution does not distinguish the component parameters between each other, then the resulting posterior distribution will be invariant with respect to permutations of the labels. Several solutions have been proposed; for a general review see Jasra, Holmes and Stephens (2005). The easiest approach is to use relabeling techniques retrospectively, by post-processing the MCMC output (Marin et al., 2005). However, in our case, we are interested in the prediction of the small area parameters, whose distribution depends on the number of areas in each latent state. Therefore, we do not use the post-processing approach and the MCMC output is permuted at every iteration according to the ordering of the mean of the response variables in each class.

5 Results

In this section we report the results of the application of the LMM area level SAE model to the LFS data presented in Section 2. We fit the model with $k = 2, \ldots, 6$ latent states. For each value of k, we run one Markov chain with 100,000 iterations and then we consider a burn-in period of 50,000 iterations. The posterior means are approximated by means of the retained MCMC samples. Similarly, the variance of the samples approximates the posterior variance of θ_{it}. We select $k = 4$ using the proposed model selection approach. In fact, using expression (A.4), we obtain the following values for the posterior density of the data: $p\left(\hat{\boldsymbol{\Theta}} \mid k=2\right)$ = 59,152.41, $p\left(\hat{\boldsymbol{\Theta}} \mid k=3\right)$ = 64,405.11, $p\left(\hat{\boldsymbol{\Theta}} \mid k=4\right)$ = 68,816.06, and $p\left(\hat{\boldsymbol{\Theta}} \mid k=5\right)$ = 68,703.75.

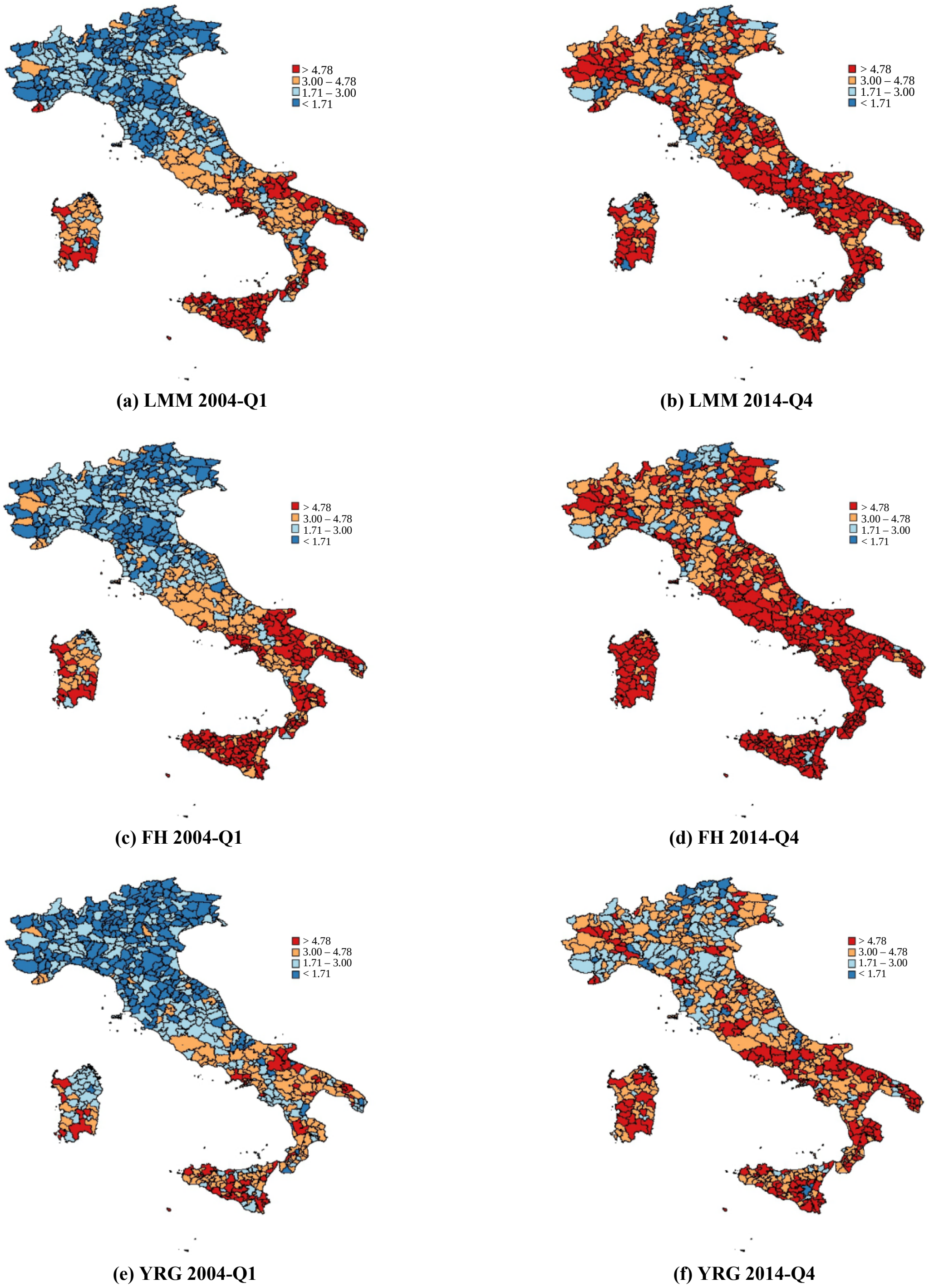
We validate our model selection procedure by comparing the final choice with that obtained using the Deviance Information Criterion (DIC). In particular, we focus on $k = 4, 5$ latent states for which the Bayes rule provides the largest values. The DIC confirms our results because we obtain 8,334.0 and 8,362.4 for $k = 4$ and $k = 5,$ respectively.

Figure 5.1 compares the map of estimates for the first and the last quarter of the whole period. These can be compared with the maps of direct estimates reported in Figure 2.1. In particular, estimates on the first row of Figure 5.1 are obtained by the proposed LMM area level model. Those on the second row are obtained using a cross-sectional Fay-Herriot (FH) model computed with the R package `hbsae` (Boonstra, 2012), while those on the last row are obtained using the You et al. (2003, YRG) model, for which we have considered three possible choices for $\rho,$ 0.50, 0.75, and 1.00, as in You et al. (2003). To measure the overall fit of the three alternative YRG models we have compared posterior predictive $p-$values (Meng, 1994). In particular, simulated values of a suitable discrepancy measure are generated from the posterior predictive distribution and, then, compared to the corresponding measure for the observed data. More specifically, if $d\left(\hat{\mathbf{\Theta}}, \mathbf{\Theta}\right)$ is a discrepancy measure that depends on the observed data, $\hat{\mathbf{\Theta}},$ and the parameter matrix $\mathbf{\Theta},$ then the posterior predictive $p-$value is defined as $P\left[d\left(\hat{\mathbf{\Theta}}^*, \mathbf{\Theta}\right) > d\left(\hat{\mathbf{\Theta}}, \mathbf{\Theta}\right) \middle| \hat{\mathbf{\Theta}}\right],$ where $\hat{\mathbf{\Theta}}^*$ is a sample from the posterior predictive distribution. If a model fits the observed data well, then the two values of the discrepancy measure are similar and, as a result, the value of the $p-$value is expected to be close to 0.5. On the other hand, $p-$values near 0 or 1 signal a model that is not well suited to the data. As in Datta et al. (1999) and in You et al. (2003), we use the following discrepancy measure

$$d\left(\hat{\mathbf{\Theta}}, \mathbf{\Theta}\right) = \sum_{i=1}^{m} \left(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i\right)' \mathbf{\Psi}_i^{-1} \left(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i\right)$$

for the overall fit. The posterior predictive measure suggests that the model with $\rho = 1$ provides a better fit to the data, in fact it takes value 0.188 for $\rho = 1.00,$ 0.103 for $\rho = 0.75,$ and 0.032 for $\rho = 0.50.$ Note that for our model, we obtain a $p-$value equal to 0.311. We have also implemented the Datta et al. (1999) estimation approach. However, the number of areas and the overall number of observations made the estimation computationally prohibitive. For this reason, it is not considered further.

From Figure 5.1, we observe that all model-based estimates are smoother than the original direct estimates. Maps are color-coded according to the quartiles of the direct estimates for 2004-Q1. In general, estimates for 2004-Q1 show a quite distinct division between North, Center, and South of Italy, with relatively higher unemployment incidences in the South of the country. For 2014-Q4, unemployment incidences are all much higher all over the country, because of the economic crisis that hit the country in 2008. LMM and FH show similar patterns, and are in line with those of the direct estimator. YRG, on the other hand, provides more shrunk estimates and this is particularly evident for 2014-Q4 where a general and distinct underestimation is provided. This behavior is displayed for all time points. In fact, Figure 5.2 shows the absolute difference between the direct estimates and model-based estimates. Areas are ordered according to estimated variance of the direct estimates. All model-based estimators show a common general behavior: smaller differences for more precise estimates and increasingly larger differences for more variable direct estimates. However, we can note that YRG provides systematically larger positive differences, by this casting some concerns on bias.



(a) LMM 2004-Q1

(b) LMM 2014-Q4

(c) FH 2004-Q1

(d) FH 2014-Q4

(e) YRG 2004-Q1

(f) YRG 2014-Q4

Figure 5.1 Unemployment incidences (%) estimated using LMM, FH and YRG for 2004-Q1 and 2014-Q4.

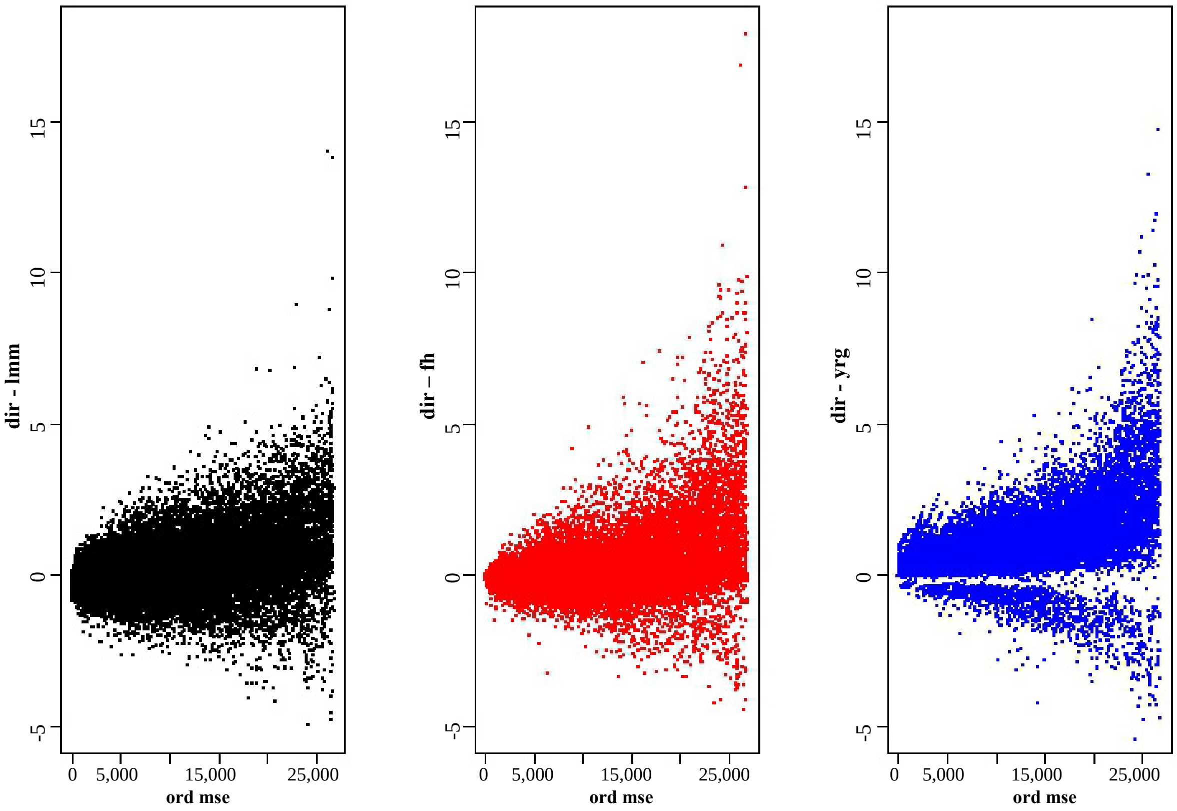


Figure 5.2 Difference between DIR and model-based small area estimates; LMM, FH, YRG, from left to right. Areas are arranged according to increasing estimated variance of the direct estimator.

As mentioned earlier, LMM uses a discrete random variable to model unobserved heterogeneity rather than the more common continuous (usually Gaussian) assumption. As a consequence, small areas can be clustered according to the latent state to which they belong at each time point. In this application, latent states are ordered and can be associated to the level of unemployment, conditionally on the covariates. Figure 5.3 shows the evolution of the latent states clustering for the small areas over the 44 time points. The fourth cluster is very small and comprises areas with a very high unemployment incidence. In addition, the pattern seems to be very stable over time, as the probability of changing latent state is very low. Note that, although there is a noticeable temporal trend in the data, this is captured by the dummy variables inserted to account for trend and seasonality. These finding are supported by the estimated initial and transition probabilities:

$$\begin{aligned}\hat{\boldsymbol{\pi}} &= (0.505,\, 0.340,\, 0.144,\, 0.011)',\\ \hat{\boldsymbol{\Pi}} &= \begin{pmatrix} 0.967 & 0.027 & 0.004 & 0.002 \\ 0.020 & 0.956 & 0.020 & 0.004 \\ 0.007 & 0.035 & 0.946 & 0.012 \\ 0.035 & 0.007 & 0.030 & 0.929 \end{pmatrix}.\end{aligned}$$

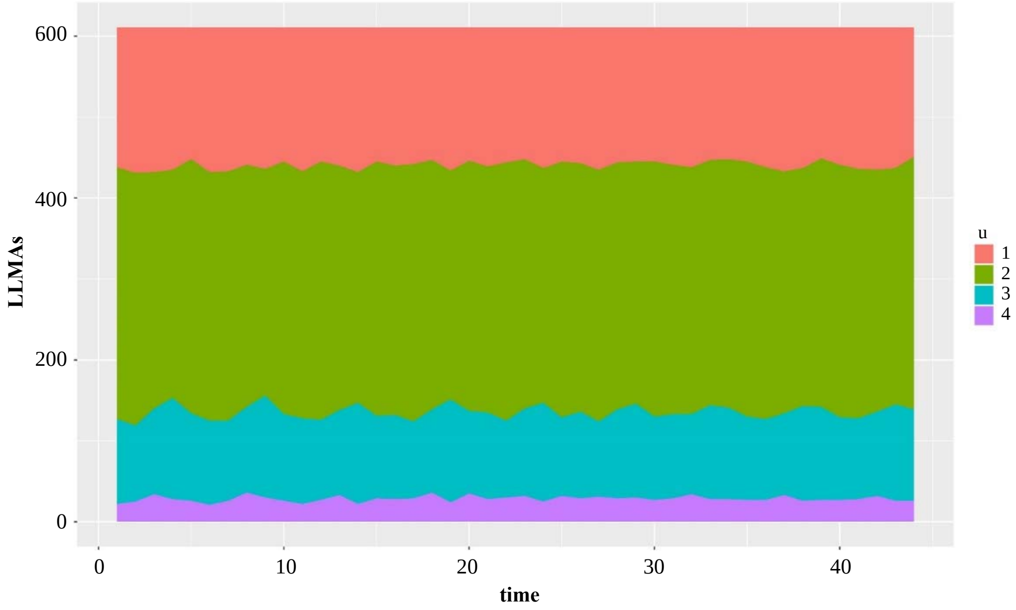
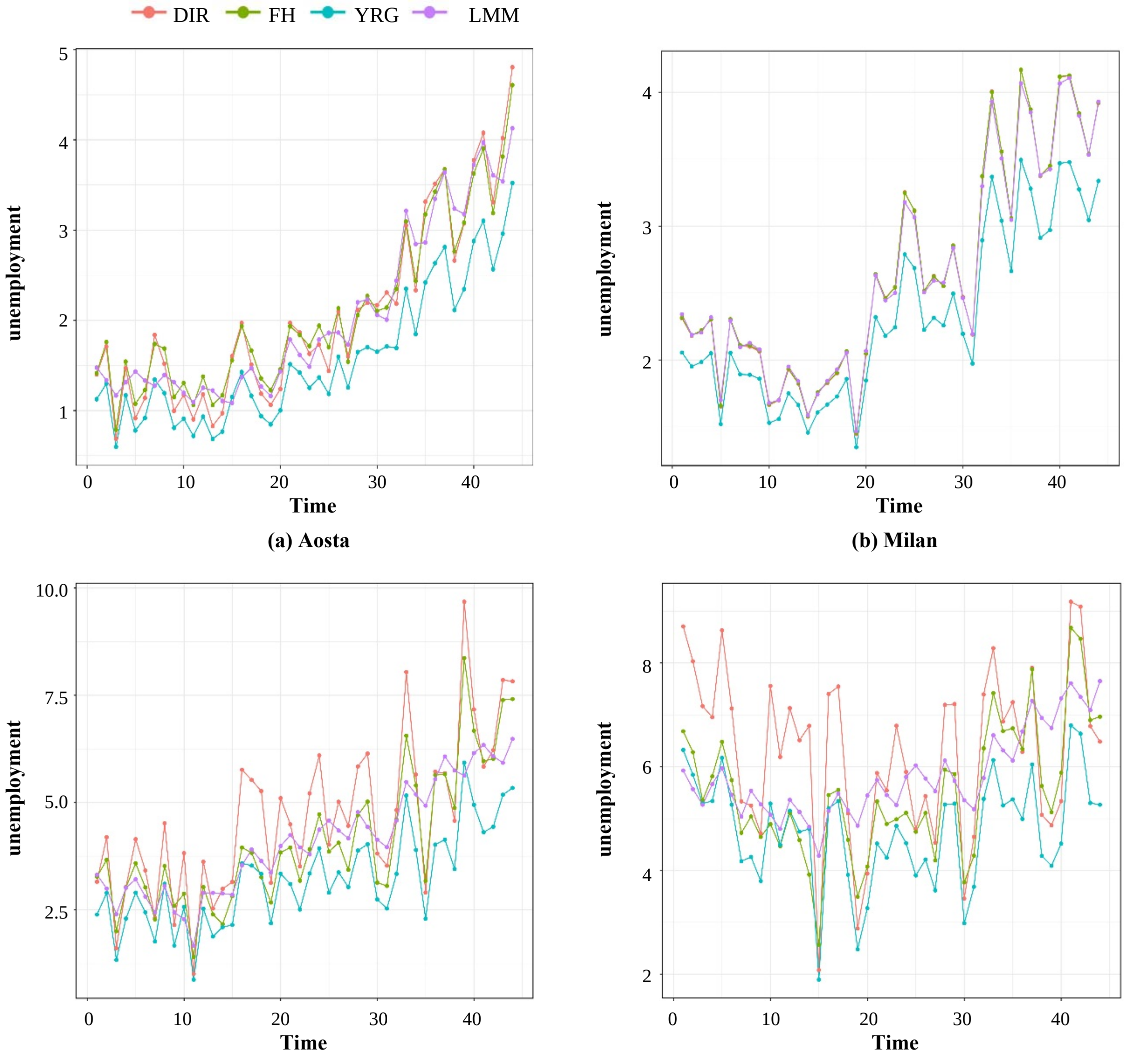


Figure 5.3 Latent states distribution from 2004-Q1 to 2014-Q4.

Figure 5.4 shows the time series of direct estimates and the corresponding model-based estimates for a selection of small areas. Aosta – panel (a) – is a small LMA in the very North of the country, with a small level of unemployment. LMM smooths the direct estimates more than the other methods, while YRG tracks the path of the direct estimates, but provides a noticeable negative bias. Milan – panel (b) – is a large city in the North of the country and the corresponding LMA has usually a very large sample size. As expected, FH and LMM track the values of DIR, while YRG exhibits a clear tendency to underestimation. Perugia and Brindisi are two mid-size towns in the Centre and in the South of Italy, respectively. The pattern of the model-based estimators is very clear: LMM provides a very good smoothing of the quite erratic trend of the direct estimates, better than FH, while YRG again displays a tendency to negative bias, particularly after the first few quarters.

It is expected that model-based estimates, besides providing estimates for the out-of-sample areas, provide gains in efficiency over direct estimates. In Figure 5.5 we report the distribution of the CV for comparing model-based small areas estimates for each time point, classified as in Figure 2.3 according to different relevant values of CV. FH provides estimates for out of sample areas, but it does not seem to provide a useful estimation option for these data since only few estimates have CV smaller than 16%. On the other hand, YRG provides a very good improvement in terms of estimated efficiency, with almost all estimates with a CV smaller than 33.3%. LMM provides a good improvement over FH with only approximately 15% of the small area estimates with a CV larger than 33.3%.



(a) Aosta

(b) Milan

(c) Perugia

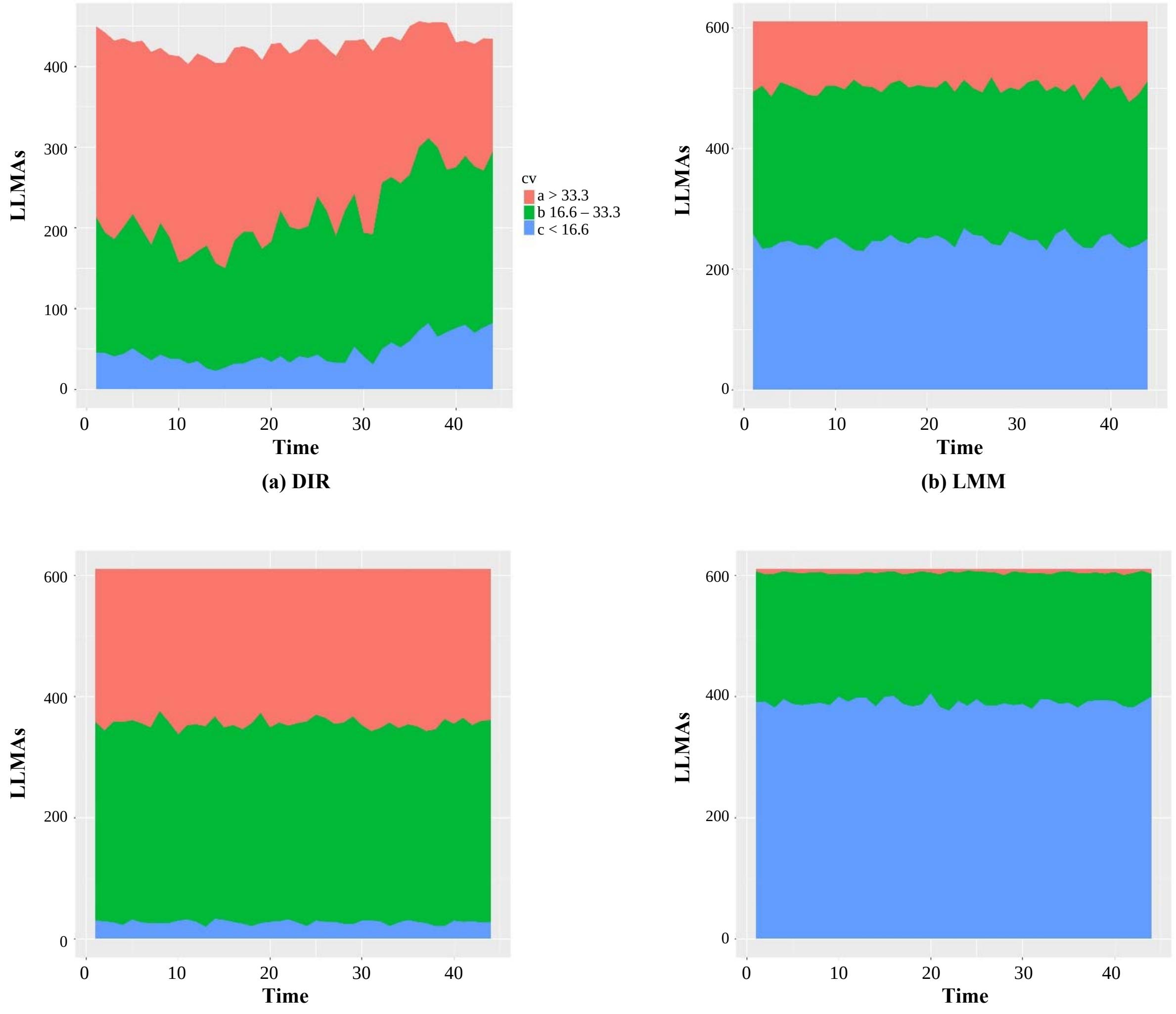
(d) Brindisi

Figure 5.4 Time series of direct and model-based estimates for a selection of four small areas.

In addition, small area estimates should be close to population level quantities, when available. Here, we use data from the 2011 Italian Population Census and consider unemployment incidence for LMAs from the Census as a gold standard. In particular, we evaluate the distance between small area estimates for the closest time point, namely 2011-Q4, and the Census value, Cens_i, and compute the Absolute Relative Error for each area (ARE_i) as

$$\text{ARE}_i = \frac{\left|\hat{\theta}_i - \text{Cens}_i\right|}{\text{Cens}_i} \tag{5.1}$$

for each area i. The ARE_i also provides a sort of measure of relative bias and is important to evaluate and compare the performance in terms of overall error of the estimates. Note that the small area parameter of interest and the Census quantity do not have exactly the same definition. In fact, the LFS is a continuous survey and the corresponding unemployment incidence refers to a quarter, while that from the Census refers to a specific calendar day. In addition, order and wording of items in the two questionnaires used to evaluate the unemployment status differ slightly. We compare the distribution of ARE_i for LMM and YRG in Figure 5.6. From the empirical distribution of ARE_i, we observe that LMM systematically provides smaller values than YRG. When looking at the subgroup of in-sample areas, we can compare this distribution with that of the direct estimator, and we conclude that LMM is in line with DIR for almost one half of the small areas, and then LMM provides estimates with a relatively smaller value of ARE_i. In conclusion, YRG estimates have a lower estimated variance, but exhibit higher estimated bias, in terms of the comparison with the Census and the direct estimates. This puts concern on coverage. On the other hand, LMM estimates are not as good as YRG estimates in terms of CV, but when looking at the bias, the overall behavior seems to be much more reliable.



(a) DIR

(b) LMM

(c) FH

(d) YRG

Figure 5.5 Distribution of the coefficients of variation for DIR, LMM, FH and YRG estimates for each quarter.

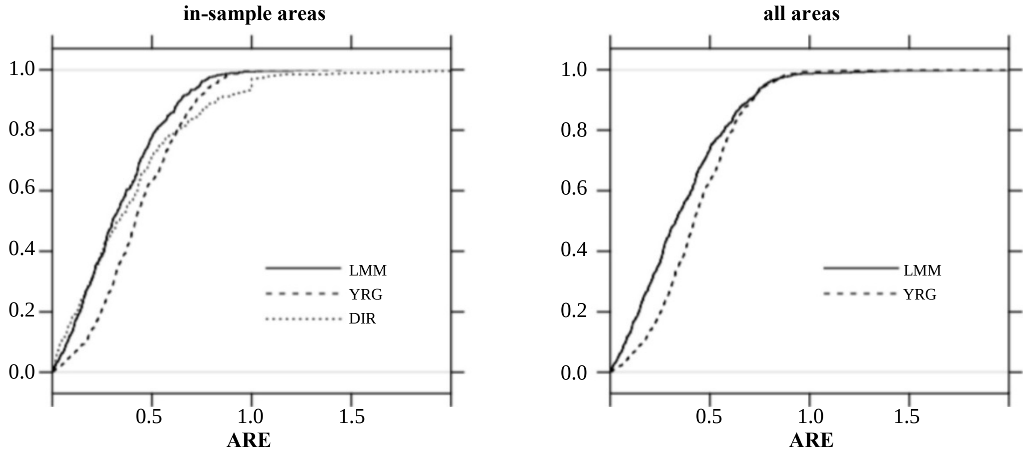


Figure 5.6 Empirical distribution of ARE_i, equation (5.1), for in-sample areas (left panel) and for all areas (right panel).

6 Final remarks

In this paper we develop a new area level SAE method that uses a Latent Markov Model (LMM) as the linking model. In LMMs (Bartolucci et al., 2013), the characteristic of interest, and its evolution in time, is represented by a latent process that follows a Markov chain, usually of first order. Under the assumption of normality for the conditional distribution of the response variables given the latent variables, the model is estimated using an augmented data Gibbs sampler. The proposed model has been applied to quarterly data from the Italian LFS from 2004 to 2014. The model-based method has been found to be effective for developing LMAs level estimates of unemployment incidence and the reduction in the coefficient of variation compared to the direct estimator is quite evident. The proposed approach is also more accurate than the direct and the time-series model-based estimator proposed by You et al. (2003) in reproducing census data. An advantage of this methodology is that it also provides a clustering of the small areas in homogeneous groups.

LMMs can be seen as an extension of latent class models to longitudinal data. In this regard, our approach represents an extension of the latent class SAE model proposed by Fabrizi et al. (2016). Moreover, LMMs may be seen as an extension of Markov chain models to control for measurement errors and can easily handle multivariate data, providing a very flexible modeling framework. The approach could be extended using spatial correlation information, and it could consider different distributions for the manifest variables, such as Poisson, Binomial, and Multinomial responses. In this scenario, we could fit unmatched sampling and linking models and handle departures from the normality assumption, but a Gibbs sampler cannot be used any longer, and Metropolis-Hastings sampling is an option. The proposed univariate model can account for measurement errors, but the extension to multivariate framework could be also possible, taking into account the conditional independence assumption.

In this application we have not accounted explicitly for the serial correlation induced by the rotating panel design. A natural way to take the different features of this design into account, such as the rotating group bias and the autocorrelation of the survey errors, is to use state space-model specifications, as in Pfeffermann (1991), Pfeffermann and Rubin-Bleuer (1993) and, more recently, Van den Brakel and Krieg (2015) and Boonstra and Van den Brakel (2016). In this context, it would also be interesting to extend to SAE the LMM with serial correlation in the measurement model proposed by Bartolucci and Farcomeni (2009). State space-model specifications can also be a useful tool to capture and model the strong trend and seasonality of this type of data.

Acknowledgements

The work of Bertarelli, Ranalli, D'Aló and Solari has been developed under the support of the project PRIN-2012F42NS8 "Household wealth and youth unemployment: new survey methods to meet current challenges". The Authors are grateful to the Associated Editor and two anonymous Referees who provided very useful comments on earlier versions of this paper.

Appendix A

Model estimation

In the following we first illustrate Bayesian estimation and model selection based on a MCMC algorithm which is implemented in a data augmentation framework (Tanner and Wong, 1987).

A.1 Data augmentation method

In order to estimate the small area parameters $\mathbf{\Theta}$, the measurement parameters $\boldsymbol{\phi}_{\text{obs}}$, and the latent parameters $\boldsymbol{\phi}_{\text{lat}}$, we follow a data augmentation approach. We recall that the observed data consist of the direct estimates $\hat{\theta}_{it}$, the corresponding smoothed $\widehat{\text{MSE}}_{it}$, and the covariate vectors $\mathbf{x}_{it}$, with $i = 1, \ldots, m$ and $t = 1, \ldots, T$. Moreover, the data augmentation approach explicitly introduces the latent variables U_{it} treated as missing data, the values of which are updated during the MCMC algorithm that is, therefore, based on a complete data likelihood. In this context, the use of conjugate priors to the complete data likelihood allows us to sample from the conditional posterior of the latent states in a straightforward way. Since the state space is finite, sampling the latent states conditionally given the model parameters is also simple.

To generate samples from the joint posterior distribution of the model parameters and latent states, the proposed MCMC algorithm proceeds as follows. Let $\hat{\mathbf{\Theta}}$ be the matrix of realizations of the available direct estimates that is defined as in (4.1), with each θ_{it} replaced by $\hat{\theta}_{it}$, and let $\mathbf{U}$ be the matrix of the latent variable U_{it}, with elements organized as in $\hat{\mathbf{\Theta}}$. Then the posterior distribution of all model parameters and latent variables, given the observed data, has the following expression:

$$p\left(\mathbf{U}, \boldsymbol{\phi}_{\text{lat}}, \boldsymbol{\phi}_{\text{obs}}, \boldsymbol{\Theta} \middle| \hat{\boldsymbol{\Theta}}\right) \propto p\left(\mathbf{U} \mid \boldsymbol{\phi}_{\text{lat}}\right) \pi\left(\boldsymbol{\phi}_{\text{lat}}\right) \pi\left(\boldsymbol{\phi}_{\text{obs}}\right) p\left(\boldsymbol{\Theta} \mid \mathbf{U}, \boldsymbol{\phi}_{\text{obs}}\right) p\left(\hat{\boldsymbol{\Theta}} \middle| \boldsymbol{\Theta}\right).$$

The MCMC algorithm alternates between sampling the latent variables and the parameters from the corresponding full conditional distribution. This scheme is repeated for R iterations. At the end of each iteration r, $r = 1, \ldots, R,$ the sampled model parameters and latent variables are obtained and are denoted by $\mathbf{U}^{(r)}, \boldsymbol{\phi}_{\text{lat}}^{(r)}, \boldsymbol{\phi}_{\text{obs}}^{(r)},$ and $\boldsymbol{\Theta}^{(r)}.$ More precisely, each iteration consists in:

1. drawing $\mathbf{U}^{(r)}$ from $p\left(\mathbf{U} \middle| \boldsymbol{\phi}_{\text{lat}}^{(r-1)}, \boldsymbol{\phi}_{\text{obs}}^{(r-1)}, \boldsymbol{\Theta}^{(r-1)}\right);$
2. drawing $\boldsymbol{\phi}_{\text{lat}}^{(r)}$ from $p\left(\boldsymbol{\phi}_{\text{lat}} \mid \mathbf{U}^{(r)}\right);$
3. drawing $\boldsymbol{\phi}_{\text{obs}}^{(r)}$ from $p\left(\boldsymbol{\phi}_{\text{obs}} \mid \mathbf{U}^{(r)}, \boldsymbol{\Theta}^{(r-1)}\right);$
4. drawing $\boldsymbol{\Theta}^{(r)}$ from $p\left(\boldsymbol{\Theta} \middle| \mathbf{U}^{(r)}, \boldsymbol{\phi}_{\text{obs}}^{(r)}, \hat{\boldsymbol{\Theta}}\right).$

In the following we illustrate in details each of the above steps. In this regard, note that our illustration is referred to the case where all elements of $\hat{\boldsymbol{\Theta}}$ are available. However, in our application, some elements of this matrix are missing. This requires minor adjustments to the MCMC algorithm, consisting in imputing the missing values by a Gibbs sampler and sampling directly from its full conditional distribution.

A.1.1 Simulation of $\mathbf{U}^{(r)}$

Each latent variable U_{it} is drawn separately from the corresponding full conditional distribution, which is of multinomial type with specific parameters. In particular, we have that

$$U_{it} \middle| U_{i,t-1}^{(r)}, U_{i,t+1}^{(r-1)}, \boldsymbol{\phi}_{\text{lat}}^{(r-1)}, \boldsymbol{\phi}_{\text{obs}}^{(r-1)}, \boldsymbol{\Theta}^{(r-1)} \sim \text{Multi}_k\left(\mathbf{q}_{it}\right), \quad t = 1, \ldots, T, \quad i = 1, \ldots, m, \qquad (A.1)$$

where $U_{i,t-1}^{(r)}$ disappears for $t = 1$ and $U_{i,t+1}^{(r)}$ disappears for $t = T.$ Moreover, the probability vector $\mathbf{q}_{it}$ is defined as follows:

- for $t = 1,$ $\mathbf{q}_{it}$ has elements proportional to

$$\pi_u^{(r-1)} \pi_{u \mid U_{i2}^{(r-1)}}^{(r-1)}, \quad u = 1, \ldots, k;$$

- for $t = 2, \ldots, T-1,$ $\mathbf{q}_{it}$ has elements proportional to

$$\pi_{u \mid U_{i,t-1}^{(r)}}^{(r-1)} \pi_{U_{i,t+1}^{(r-1)} \mid u}^{(r-1)}, \quad u = 1, \ldots, k;$$

- for $t = T,$ $\mathbf{q}_{it}$ has elements proportional to

$$\pi_{u \mid U_{i,T-1}^{(r)}}^{(r-1)}, \quad u = 1, \ldots, k.$$

A.1.2 Simulation of $\boldsymbol{\phi}_{\text{lat}}^{(r)}$

Recalling that $\boldsymbol{\phi}_{\text{lat}} = \{\boldsymbol{\pi}, \boldsymbol{\Pi}\},$ we first draw $\boldsymbol{\pi}^{(r)}$ from the full conditional distribution:

$$\boldsymbol{\pi} \,|\, \mathbf{U}^{(r)} \sim \text{Dirichlet}\left(\mathbf{1}_k + \mathbf{n}_1\right),$$

where $\mathbf{n}_1 = \left(n_{11}, \ldots, n_{1k}\right)'$ and n_{1u} is the number of areas in state u at time 1, with $u = 1, \ldots, k$. Moreover, we draw each row of matrix $\mathbf{\Pi}$ from the distribution

$$\boldsymbol{\pi}_{\bar{u}} \,|\, \mathbf{U}^{(r)} \sim \text{Dirichlet}\left(\mathbf{1}_k + \mathbf{n}_{\bar{u},t}\right), \qquad t = 2, \ldots, T,$$

where $\mathbf{n}_{\bar{u},t} = \left(n_{\bar{u},t_1}, \ldots, n_{\bar{u},t_k}\right)'$ and $n_{\bar{u},t_u}$ is the number of areas moving from state $\bar{u}$ to state u at time t, with $t = 2, \ldots, T$ and $u, \bar{u} = 1, \ldots, k$.

A.1.3 Simulation of $\boldsymbol{\phi}_{\text{obs}}^{(r)}$

Considering that $\boldsymbol{\phi}_{\text{obs}} = \{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k, \sigma_1^2, \ldots, \sigma_k^2\}$, we first draw each $\boldsymbol{\beta}_u$, $u = 1, \ldots, k$, from the full conditional distribution:

$$\boldsymbol{\beta}_u \,|\, \mathbf{U}^{(r)}, \mathbf{\Theta}^{(r)} \sim N_p\left(\boldsymbol{\eta}_{1,u}, \mathbf{\Sigma}_{1,u}\right),$$

where

$$\begin{aligned}
\boldsymbol{\eta}_{1,u} &= \mathbf{\Lambda}_{1,u}^{-1} \sum_{i=1}^{m} \sum_{t=1}^{T} \theta_{it} I\left(U_{it} = u\right) \mathbf{x}_{it}, \\
\mathbf{\Sigma}_{1,u} &= \sigma_u^2 \mathbf{\Lambda}_{1,u}^{-1}, \\
\mathbf{\Lambda}_{1,u} &= \sum_{i=1}^{m} \sum_{t=1}^{T} \mathbf{x}_{it} \mathbf{x}_{it}' I\left(U_{it} = u\right) + \mathbf{\Lambda}_0,
\end{aligned}$$

with $I(\cdot)$ denoting the indicator function equal to 1 if its argument is true and to 0 otherwise. Then, we draw each σ_u^2 from

$$\sigma_u^2 \,|\, \mathbf{U}^{(r)}, \mathbf{\Theta}^{(r)} \sim \text{IG}\left(a_{1,u}, b_{1,u}\right),$$

with

$$\begin{aligned}
a_{1,u} &= a_0 + \frac{n_{.u}}{2}, \\
b_{1,u} &= b_0 + \frac{1}{2}\left(\sum_{i=1}^{m} \sum_{t=1}^{T} \theta_{it}^2 I\left(U_{it} = u\right) + \boldsymbol{\eta}_0' \mathbf{\Lambda}_0 \boldsymbol{\eta}_0 - \boldsymbol{\eta}_{1,u}' \mathbf{\Lambda}_{1,u} \boldsymbol{\eta}_{1,u}\right),
\end{aligned}$$

where $n_{.u} = \sum_{t=1}^{T} n_{tu}$ is the number of areas in state u regardless of the specific time occasion.

A.1.4 Simulation of $\mathbf{\Theta}^{(r)}$

The goal of SAE is to predict each θ_{it}, $i = 1, \ldots, m$, $t = 1, \ldots, T$, based on the model and the observed data. This amounts to draw these elements from

$$\theta_{it} \,|\, \mathbf{U}^{(r)}, \boldsymbol{\phi}_{\text{obs}}^{(r)}, \hat{\theta}_{it} \sim N(\hat{\theta}_{it}^{(r)}, \gamma_{it}^{(r)} \psi_{it}),$$

where

$$\hat{\theta}_{it}^{(r)} = \gamma_{it}^{(r)}\hat{\theta}_{it} + \left(1-\gamma_{it}^{(r)}\right)\mathbf{x}'_{it}\boldsymbol{\beta}_u^{(r)}, \tag{A.2}$$

with $\gamma_{it}^{(r)} = \sigma_u^{2(r)} \Big/ \left(\sigma_u^{2(r)} + \psi_{it}\right)$.

A.2 Model selection: The Chib estimator

The method proposed in Chib (1995) can be applied to perform model selection starting from the Gibbs sampler output. It is known that the posterior density can be written as the ratio of the product of the likelihood function and the priors divided by the marginal likelihood:

$$p\left(\mathbf{U}, \boldsymbol{\phi}_{\text{lat}}, \boldsymbol{\phi}_{\text{obs}}, \boldsymbol{\Theta} \,\middle|\, \hat{\boldsymbol{\Theta}}\right) = \frac{p\,(\mathbf{U} \,|\, \boldsymbol{\phi}_{\text{lat}})\,\pi\,(\boldsymbol{\phi}_{\text{lat}})\,\pi\,(\boldsymbol{\phi}_{\text{obs}})\,p\,(\boldsymbol{\Theta} \,|\, \mathbf{U}, \boldsymbol{\phi}_{\text{obs}})\,p\left(\hat{\boldsymbol{\Theta}} \,|\, \boldsymbol{\Theta}\right)}{p\left(\hat{\boldsymbol{\Theta}}\right)}. \tag{A.3}$$

Therefore, it is possible to write the marginal likelihood of the data $\hat{\boldsymbol{\Theta}}$ as

$$p\left(\hat{\boldsymbol{\Theta}}\right) = \frac{p\left(\hat{\boldsymbol{\Theta}} \,|\, \boldsymbol{\Theta}\right) p\,(\mathbf{U} \,|\, \boldsymbol{\phi}_{\text{lat}})\,\pi\,(\boldsymbol{\phi}_{\text{lat}})\,\pi\,(\boldsymbol{\phi}_{\text{obs}})\,p\,(\boldsymbol{\Theta} \,|\, \mathbf{U}, \boldsymbol{\phi}_{\text{obs}})}{p\left(\mathbf{U}, \boldsymbol{\phi}_{\text{lat}}, \boldsymbol{\phi}_{\text{obs}}, \boldsymbol{\Theta} \,\middle|\, \hat{\boldsymbol{\Theta}}\right)}, \tag{A.4}$$

for any $\mathbf{U}, \boldsymbol{\phi}_{\text{lat}}, \boldsymbol{\phi}_{\text{obs}}, \boldsymbol{\Theta}$ and $\hat{\boldsymbol{\Theta}}$. We drop the dependence on k for ease of notation. This is the model selection criterion used in Section 5. Then, choosing specific values of the latent variables and model parameters, denoted by $\bar{\mathbf{U}}, \bar{\boldsymbol{\phi}}_{\text{lat}}, \bar{\boldsymbol{\phi}}_{\text{obs}},$ and $\bar{\boldsymbol{\Theta}},$ we can estimate $\log p\left(\hat{\boldsymbol{\Theta}}\right)$ through the following decomposition:

$$\begin{aligned}\log p\left(\hat{\boldsymbol{\Theta}}\right) = {} & \log p\left(\hat{\boldsymbol{\Theta}} \,\middle|\, \bar{\boldsymbol{\Theta}}\right) + \log p\left(\bar{\mathbf{U}} \,\middle|\, \bar{\boldsymbol{\phi}}_{\text{lat}}\right) + \log \pi\left(\bar{\boldsymbol{\phi}}_{\text{lat}}\right) + \log \pi\left(\bar{\boldsymbol{\phi}}_{\text{obs}}\right) \\ & + \log p\left(\bar{\boldsymbol{\Theta}} \,\middle|\, \bar{\mathbf{U}}, \bar{\boldsymbol{\phi}}_{\text{obs}}\right) - \log p\left(\bar{\mathbf{U}}, \bar{\boldsymbol{\phi}}_{\text{lat}}, \bar{\boldsymbol{\phi}}_{\text{obs}}, \bar{\boldsymbol{\Theta}} \,\middle|\, \hat{\boldsymbol{\Theta}}\right).\end{aligned} \tag{A.5}$$

The use of the log transformation is motivated by numerical stability (Chib, 1995).

The first five terms at the right hand side of (A.5) can be computed directly from the assumed distributions of the parameters and the data. On the other hand obtaining the last component is more challanging. By the law of total probability, $p\left(\bar{\mathbf{U}}, \bar{\boldsymbol{\phi}}_{\text{lat}}, \bar{\boldsymbol{\phi}}_{\text{obs}}, \bar{\boldsymbol{\Theta}} \,\middle|\, \hat{\boldsymbol{\Theta}}\right)$ may be decomposed as

$$p\left(\bar{\mathbf{U}}, \bar{\boldsymbol{\phi}}_{\text{lat}}, \bar{\boldsymbol{\phi}}_{\text{obs}}, \bar{\boldsymbol{\Theta}} \,\middle|\, \hat{\boldsymbol{\Theta}}\right) = p\left(\bar{\mathbf{U}} \,\middle|\, \bar{\boldsymbol{\phi}}_{\text{lat}}, \bar{\boldsymbol{\phi}}_{\text{obs}}, \bar{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Theta}}\right) p\left(\bar{\boldsymbol{\phi}}_{\text{lat}} \,\middle|\, \bar{\boldsymbol{\phi}}_{\text{obs}}, \bar{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Theta}}\right) p\left(\bar{\boldsymbol{\phi}}_{\text{obs}} \,\middle|\, \bar{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Theta}}\right) p\left(\bar{\boldsymbol{\Theta}} \,\middle|\, \hat{\boldsymbol{\Theta}}\right). \tag{A.6}$$

Following Chib (1995), we compute the first term of (A.6) following the Gibbs scheme outlined in Section A.1, whereas, the other three terms are estimated from the Gibbs output. In particular, we estimate

$$p\left(\bar{\boldsymbol{\phi}}_{\text{lat}} \,\middle|\, \bar{\boldsymbol{\phi}}_{\text{obs}}, \bar{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Theta}}\right) = \int p\left(\bar{\boldsymbol{\phi}}_{\text{lat}} \,\middle|\, \bar{\mathbf{U}}, \bar{\boldsymbol{\phi}}_{\text{obs}}, \bar{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Theta}}\right) p\left(\bar{\mathbf{U}} \,\middle|\, \bar{\boldsymbol{\phi}}_{\text{obs}}, \bar{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Theta}}\right) d\bar{\mathbf{U}}$$

as $R^{-1}\sum_{r=1}^{R} p\left(\bar{\boldsymbol{\phi}}_{\text{lat}} \,|\, \mathbf{U}^{(r)}, \bar{\boldsymbol{\phi}}_{\text{obs}}, \bar{\boldsymbol{\Theta}}\right)$, based on R draws from a reduced Gibbs sampling where $\mathbf{U}$ is not updated. In order to estimate

$$p\left(\bar{\boldsymbol{\phi}}_{\text{obs}} \,\middle|\, \bar{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Theta}}\right) = \int p\left(\bar{\boldsymbol{\phi}}_{\text{obs}} \,\middle|\, \bar{\mathbf{U}}, \bar{\boldsymbol{\phi}}_{\text{lat}}, \bar{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Theta}}\right) p\left(\bar{\mathbf{U}}, \bar{\boldsymbol{\phi}}_{\text{lat}} \,\middle|\, \bar{\boldsymbol{\Theta}}, \hat{\boldsymbol{\Theta}}\right) d\bar{\mathbf{U}}\, d\bar{\boldsymbol{\phi}}_{\text{lat}},$$

we use $R^{-1}\sum_{r=1}^{R} p\left(\bar{\boldsymbol{\phi}}_{\text{obs}} \mid \mathbf{U}^{(r,1)}, \bar{\boldsymbol{\phi}}_{\text{lat}}^{(r,1)}, \bar{\boldsymbol{\Theta}}\right)$. Finally, to estimate

$$p\left(\bar{\boldsymbol{\Theta}} \middle| \hat{\boldsymbol{\Theta}}\right) = \int p\left(\bar{\boldsymbol{\Theta}} \middle| \bar{\mathbf{U}}, \bar{\boldsymbol{\phi}}_{\text{lat}}, \bar{\boldsymbol{\phi}}_{\text{obs}}, \hat{\boldsymbol{\Theta}}\right) p\left(\bar{\mathbf{U}}, \bar{\boldsymbol{\phi}}_{\text{lat}}, \bar{\boldsymbol{\phi}}_{\text{obs}} \middle| \hat{\boldsymbol{\Theta}}\right) d\bar{\mathbf{U}}\, d\bar{\boldsymbol{\phi}}_{\text{lat}}\, d\bar{\boldsymbol{\phi}}_{\text{obs}},$$

we use $R^{-1}\sum_{r=1}^{R} p\left(\bar{\boldsymbol{\Theta}} \mid \mathbf{U}^{(r,\,2)}, \boldsymbol{\phi}_{\text{lat}}^{(r,\,2)}, \boldsymbol{\phi}_{\text{obs}}^{(r,\,2)}\right)$, with R draws from a third reduced Gibbs sampling.

References

Bartolucci, F., and Farcomeni, A. (2009). A multivariate extension of the dynamic logit model for longitudinal data based on a latent Markov heterogeneity structure. *Journal of the American Statistical Association*, 104, 816-831.

Bartolucci, F., Farcomeni, A. and Pennoni, F. (2013). *Latent Markov Models for Longitudinal Data*. Boca Roton, FL: CRC Press.

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. *The Annals of Applied Statistics*, 3, 611-636.

Bartolucci, F., Pennoni, F. and Francis, B. (2007). A latent Markov model for detecting patterns of criminal activity. *Journal of the Royal Statistical Society, Series A*, 170, 115-132.

Boonstra, H.J. (2012). hbsae: Hierarchical Bayesian small area estimation. *R Package Version 1*.

Boonstra, H.J. (2014). Time-series small area estimation for unemployment based on a rotating panel survey. Technical report, CBS. Available at https://www.cbs.nl/nl-nl/achtergrond/2014/25/time-series-small-area-estimation-for-unemployment-based-on-a-rotating-panel-survey.

Boonstra, H.J., and van den Brakel, J.A. (2016). *Estimation of Level and Change for Unemployment Using Multilevel and Structural Time Series Models*. Discussion paper 2016-10. Statistics Netherlands, Heerlen.

Boys, R., and Henderson, D. (2003). Data augmentation and marginal updating schemes for inference in hidden Markov models. Technical report, Univ. Newcastle.

Carlin, B.P., and Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B*, 56, 473-484.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90, 1313-1321.

D’Alò, M., Di Consiglio, L., Falorsi, S., Ranalli, M.G. and Solari, F. (2012). Use of spatial information in small area models for unemployment rate estimation at sub-provincial areas in Italy. *Journal of the Indian Society of Agricultural Statistics*, 66, 43-53.

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

Fabrizi, E., Montanari, G.E. and Ranalli, M.G. (2016). A hierarchical latent class model for predicting disability small area counts from survey data. *Journal of the Royal Statistical Society, Series A*, 179, 103-132.

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, 269-277.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3), 515-534.

Gelman, A., Jakulin, A., Pittau, M.G. and Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics,* 2, 1360-1383.

Germain, S.E. (2010). *Bayesian Spatio-Temporal Modelling of Rainfall Through Non-Homogenous Hidden Markov Models*. Ph.D. thesis, University of Newcastle Upon Tyne.

Ghosh, M., Nangia, N. and Kim, D.H. (1996). Estimation of median income of four-person families: A Bayesian time series approach. *Journal of the American Statistical Association*, 91, 1423-1431.

Harvey, A., and Chung, C.-H. (2000). Estimating the underlying change in unemployment in the UK. *Journal of the Royal Statistical Society, Series A*, 163, 303-309.

Istat (2014). I sistemi locali del lavoro 2011. *Rapporto Annuale 2014*.

Jasra, A., Holmes, C. and Stephens, D. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science,* 20, 50-67.

Krieg, S., and van der Brakel, J.A. (2012). Estimation of the monthly unemployment rate for six domains through structural time series modelling with cointegrated trends. *Computational Statistics & Data Analysis,* 56, 2918-2933.

Lazarsfeld, P.F., Henry, N.W. and Anderson, T.W. (1968). *Latent Structure Analysis*. Houghton Mifflin Boston.

Liu, J.S., Wong, W.H. and Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, 81, 27-40.

MacDonald, I.L., and Zucchini, W. (1997). *Hidden Markov and Other Models for Discrete-Valued Time Series*. London: Series Chapman & Hall.

Marhuenda, Y., Molina, I. and Morales, D. (2013). Small area estimation with spatio-temporal Fay-Herriot models. *Computational Statistics & Data Analysis,* 58, 308-325.

Marin, J.-M., Mengersen, K. and Robert, C.P. (2005). Bayesian modelling and inference on mixtures of distributions. *Handbook of Statistics,* 25, 459-507.

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

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

Pfeffermann, D., and Burck, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology*, 16, 2, 217-237. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/1990002/article/14534-eng.pdf.

Pfeffermann, D., and Rubin-Bleuer, S. (1993). Robust joint modelling of labour force series of small areas. *Survey Methodology*, 19, 2, 149-163. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/1993002/article/14458-eng.pdf.

Pfeffermann, D., and Tiller, R. (2006). Small-area estimation with state-space models subject to benchmark constraints. *Journal of the American Statistical Association*, 101, 1387-1397.

Polson, N.G., and Scott, J.G. (2012). On the half-cauchy prior for a global scale parameter. *Bayesian Analysis*, 7, 4, 887-902.

Rao, J.N.K. (2003). *Small Area Estimation*. Wiley Online Library.

Rao, J.N.K., and Yu, M. (1994). Small-area estimation by combining time-series and cross-sectional data. *The Canadian Journal of Statistics,* 22, 4, 511-528.

Spezia, L. (2010). Bayesian analysis of multivariate gaussian hidden Markov models with an unknown number of regimes. *Journal of Time Series Analysis*, 31, 1-11.

Statistics Canada (2016). Guide to the Labour Force Survey. Technical report, Statistics Canada, Catalogue 71-543-G, available at http://www.statcan.gc.ca/pub/71-543-g/71-543-g2016001-eng.pdf.

Tanner, M.A., and Wong, W.H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association,* 82, 528-540.

Van den Brakel, J.A., and Krieg, S. (2015). Dealing with small sample sizes, rotation group bias and discontinuities in a rotating panel design. *Survey Methodology*, 41, 2, 267-296. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2015002/article/14231-eng.pdf.

Van der Brakel, J.A., and Krieg, S. (2016). Small area estimation with state space common factor models for rotating panels. *Journal of the Royal Statistical Society, Series A*, 179, 763-791.

Van Dyk, D.A., and Meng, X.-L. (2001). The art of data augmentation. *Journal of Computational and Graphical Statistics,* 10(1), 1-50.

Vermunt, J.K., and Magidson, J. (2002). Latent class cluster analysis. *Applied Latent Class Analysis*, 11, 89-106.

Wiggins, L.M. (1973). *Panel Analysis: Latent Probability Models for Attitude and Behavior Processes*. Jossey-Bass.

Wolter, K. (2007). *Introduction to Variance Estimation*. New York: Springer Science & Business Media.

You, Y., Rao, J.N.K. and Gambino, J. (2003). Model-based unemployment rate estimation for the Canadian Labour Force Survey: A hierarchical Bayes approach. *Survey Methodology*, 29, 1, 25-32. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2003001/article/6602-eng.pdf.

Survey Methodology, December 2018
Vol. 44, No. 2, pp. 193-214
Statistics Canada, Catalogue No. 12-001-X

Sample-based estimation of mean electricity consumption curves for small domains

Anne De Moliner and Camelia Goga[1]

Abstract

Many studies conducted by various electric utilities around the world are based on the analysis of mean electricity consumption curves for various subpopulations, particularly geographic in nature. Those mean curves are estimated from samples of thousands of curves measured at very short intervals over long periods. Estimation for small subpopulations, also called small domains, is a very timely topic in sampling theory.

In this article, we will examine this problem based on functional data and we will try to estimate the mean curves for small domains. For this, we propose four methods: functional linear regression; modelling the scores of a principal component analysis by unit-level linear mixed models; and two non-parametric estimators, with one based on regression trees and the other on random forests, adapted to the curves. All these methods have been tested and compared using real electricity consumption data for households in France.

Key Words: Regression trees; functional data; random forests; linear mixed models; robustness.

1 Introduction and context

Many studies conducted by the French electric company EDF are based on the analysis of the mean curves of electricity consumption by groups of customers who share common characteristics (e.g., similar electrical equipment or a common rate). In this text, these groups will be called domains. These mean consumption curves, also called demand curves, are estimated using a sample of several thousand curves measured at half-hourly intervals over long periods (often years).

In the literature, estimation of a total or mean demand curve for various sampling plans and the construction of confidence intervals has been examined in the recent work of Cardot, Dessertaine, Goga, Josserand and Lardin (2013), Cardot, Degras and Josserand (2013), and Cardot, Goga and Lardin (2013). The estimation of totals or means for functional data raises specific problems regarding the sample estimate of the finite population, as the strong time dependencies of the data must be exploited and preserved.

Here, we will focus on the problem of estimating mean curves for small domains, i.e., cases where we look simultaneously at several subpopulations, which may be small in size. With the advent of smart meters, it will become increasingly easy and less and less costly to create and maintain large samples of demand curves. It will therefore be possible to produce estimates of mean curves not only throughout France, but also for small geographic areas such as regions, departments and even cities. For example, these estimates could be used to propose services based on an analysis of consumption curves in territorial communities or for publication as part of an open data process.

This issue of small domains is frequently addressed in sampling theory outside the framework of functional data. The recent book by Rao and Molina (2015) proposes a state-of-the-art report on existing

[1] Anne De Moliner, IMB, Université de Bourgogne Franche-Comté / EDF R&D Paris-Saclay. E-mail: anne.de-moliner@enedis.fr; Camelia Goga, LMB, Université de Bourgogne Franche-Comté. E-mail: camelia.goga@univ-fcomte.fr.

methods. When the domain is small, direct estimators (i.e., constructed solely from individuals in the sample within the domain) are not very effective. To improve the quality of estimates, auxiliary information is used and estimators are constructed based on implicit or explicit modelling of the link between quantity of interest and auxiliary information, common to all domains. In the context of EDF, this auxiliary information can, for example, be from known billing data (rate, contract power, total consumption in the previous year in particular) for each individual in the population, but also from open data proposed by the INSEE for small geographic aggregates (IRIS).

In the literature, there are estimation methods for small domains specific to temporal series. For example, Pfefferman and Burck (1990) and Rao and Yu (1994) superimpose temporal series models on series of variables and/or coefficients of the various instants to take into consideration temporal dependencies. However, those space-state-type models were developed for relatively short temporal series (a few dozen points). They are estimated using Kalman filters, which require a lot of calculation time, which would present a problem in our context, in which the number of domains studied can vary widely.

To our knowledge, the estimation of small domains in surveys for functional data has not yet been examined in the literature. To address this problem, we propose two types of methods. First, we apply parametric methods such as linear mixed models and functional linear regressions to the coordinates of the projected curves in a finite base, e.g., the base of principal components of a principal components analysis. We also propose two non-parametric methods based on regression trees and random forests adapted to the curves, respectively. All these methods are part of the model-based survey approach.

In Section 2, we formalize the problem and introduce a few notations. In Section 3, we present two direct estimators (the Horvitz-Thompson estimator and the calibration estimator for functional data) that will be the references to which we will compare ourselves to evaluate the performance of our methods. In Section 4, we propose two parametric methods based on unit-level linear functional and mixed models, adapted to the context of the functional data, and two non-parametric methods based on regression trees and random forests. For each method, we also propose a procedure for approximating the bootstrap variance. Finally, in Section 5, all estimation methods proposed in this article are tested and compared to a data set from actual electricity consumption curves for households in France. The conclusions and perspectives are presented in Section 6. In particular, the respective benefits and drawbacks of the various methods are compared in Subsection 5.4.

2 Notations and framework

Let a population of interest U of size N. A (demand) curve $Y_i(t)$ measured for each instant t belonging to an interval of time $[0, T]$ is associated with each unit i of the population. The population U can be decomposed in D disjoint domains U_d of known sizes N_d, $d = 1, \ldots, D$. Our goal is to estimate the mean curve of Y in each domain:

$$\mu_d(t) = \frac{1}{N_d} \sum_{i \in U_d} Y_i(t), \quad t \in [0, T], \quad d = 1, \ldots, D. \tag{2.1}$$

In the population $U,$ we select a sample s of size $n,$ based on a random sampling design $p(\cdot).$ Let $\pi_i = \Pr(i \in s)$ the probability of inclusion of unit i in sample s and assumed to be positive for all units $i \in U.$ Let $s_d = s \cap U_d$ the portion of s belonging to domain U_d of random size $n_d,$ which can be equal to 0 for one or more domains.

We assume that we also have a dimensional vector p of auxiliary variables (non-dependent on time) $\mathbf{X}_i$ that will be assumed to be known for each individual i in the population and with a known average $\overline{\mathbf{X}}_d = \sum_{i \in U_d} \mathbf{X}_i / N_d$ on the domain $d = 1, \ldots, D.$

In practice, the curves are not observed continuously, but only for a series of measurement instants $0 = t_1 < t_2 < \ldots < t_L = T$ that are also assumed to be equidistant and identical for all individuals. It is also assumed that there are no missing values.

3 Direct estimation methods in the design-based approach

In this section, we adopt the sampling design approach. This means that the variable interest values Y_i for each population unit are considered to be deterministic and the only variable present is that of the construction of the sample. The statistical inference then only describes the randomness created by the sampling design.

We will present two classic estimators, the Horvitz-Thompson estimator and the calibration estimator, which will be the references to which we will compare our methods to evaluate performances. These are direct estimators, i.e., estimators constructed by using, for the estimation of the mean for each domain, only units and auxiliary information related to the domain in question.

The functional Horvitz-Thompson estimator (Horvitz and Thompson, 1952; Cardo, Chaouch, Goga and Labruère, 2010) of μ_d is given by:

$$\hat{\mu}_d^{\mathrm{HT}}(t) = \frac{1}{N_d} \sum_{i \in s_d} d_i Y_i(t), \qquad d = 1, \ldots, D, \qquad t \in [0, T], \tag{3.1}$$

with $d_i = 1/\pi_i$ the sampling weight of unit $i,$ also called the Horvitz-Thompson weight. It obviously cannot be calculated for the unsampled domains (i.e., domains d such that s_d is empty) and it is extremely unstable for small domains. Moreover, it in no way uses the predictor variables available to us.

To take advantage of the auxiliary information, again in a sampling design approach, we can use the calibration estimator proposed by Deville and Särndal (1992).

The calibration estimator for the mean μ_d is given by:

$$\hat{\mu}_d^{\mathrm{cal}}(t) = \frac{1}{N_d} \sum_{i \in s_d} w_{id}^{\mathrm{cal}} Y_i(t) \qquad d = 1, \ldots, D, \qquad t \in [0, T], \tag{3.2}$$

where the calibration weights $w_{id}^{\mathrm{cal}}, i \in s_d$ are as close as possible to the sampling weights d_i units of s_d within the meaning of a certain distance or pseudo-distance $G(w, d)$ defined by the statistician:

$$\min_{w_{id}} \sum_{i \in s_d} d_i G(w_{id}, d_i) \quad \text{subject to} \quad \sum_{i \in s_d} w_{id} \mathbf{X}_i = \sum_{i \in U_d} \mathbf{X}_i. \tag{3.3}$$

For the distance of chi-square $G(w_{id}, d_i) = \sum_{i \in s_d} (w_{id} - d_i)^2 \big/ d_i$, the weights are given by

$$w_{id}^{\text{cal}} = d_i + d_i \left(\sum_{i \in U_d} \mathbf{X}_i - \sum_{i \in s_d} d_i \mathbf{X}_i \right)' \left(\sum_{i \in s_d} d_i \mathbf{X}_i \mathbf{X}_i' \right)^{-1} \mathbf{X}_i, \quad i \in s_d$$

and the estimator becomes

$$\hat{\mu}_d^{\text{cal}}(t) = \frac{1}{N_d} \sum_{i \in s_d} d_i y_i - \frac{1}{N_d} \left(\sum_{i \in s_d} d_i \mathbf{X}_i - \sum_{i \in U_d} \mathbf{X}_i \right)' \hat{\boldsymbol{\beta}}_d(t),$$

where $\hat{\boldsymbol{\beta}}_d(t) = \left(\sum_{i \in s_d} d_i \mathbf{X}_i \mathbf{X}_i' \right)^{-1} \sum_{i \in s_d} d_i \mathbf{X}_i Y_i(t)$. The calibration weights are not dependent on time t, but they are dependent in this case on the domain d, therefore, the estimator $\hat{\mu}_d^{\text{cal}}(t)$ does not satisfy the additivity property, i.e., $\sum_{d=1}^{D} \hat{\mu}_d^{\text{cal}}(t) = \hat{\mu}^{\text{cal}}(t)$ where $\hat{\mu}^{\text{cal}}(t)$ is the calibration estimator of $\mu = \sum_{i \in U} Y_i \big/ N$. Where the vector $\mathbf{1} = (1, 1, \ldots, 1)'$ is in the model, thus,

$$\hat{\mu}_d^{\text{cal}}(t) = \frac{1}{N_d} \sum_{i \in U_d} \mathbf{X}_i' \hat{\boldsymbol{\beta}}_d(t) = \overline{\mathbf{X}}_d \hat{\boldsymbol{\beta}}_d(t), \quad t \in [0, T].$$

If size n_d is large, this estimator is approximately bias-free regarding the sampling plan. We can consider the modified estimator:

$$\hat{\mu}_d^{\text{mod}}(t) = \frac{1}{N_d} \sum_{i \in s_d} d_i Y_i(t) - \frac{1}{N_d} \left(\sum_{i \in s_d} d_i \mathbf{X}_i - \sum_{i \in U_d} \mathbf{X}_i \right)' \hat{\boldsymbol{\beta}}(t), \quad t \in [0, T], \tag{3.4}$$

where

$$\hat{\boldsymbol{\beta}}(t) = \left(\sum_{i \in s} d_i \mathbf{X}_i \mathbf{X}_i' \right)^{-1} \sum_{i \in s} d_i \mathbf{X}_i Y_i(t), \quad t \in [0, T], \tag{3.5}$$

does not depend on domain d and, therefore, the estimator $\hat{\mu}_d^{\text{mod}}$ satisfies the additivity property, i.e., $\sum_{d=1}^{D} \hat{\mu}_d^{\text{mod}}(t) = \hat{\mu}^{\text{cal}}(t)$ where $\hat{\mu}^{\text{cal}}(t)$ is the calibration estimator of $\mu = \sum_{i \in U} Y_i \big/ N$. As well, if n is large, it has no asymptotic bias even if size n_d is not large. The asymptotic variance functions of $\hat{\mu}_d^{\text{cal}}(t)$ and $\hat{\mu}_d^{\text{mod}}(t)$ are equal to the Horvitz-Thompson variances of residuals $Y_i(t) - \mathbf{X}_i' \hat{\boldsymbol{\beta}}_d(t)$ and $Y_i(t) - \mathbf{X}_i' \hat{\boldsymbol{\beta}}(t)$ (see Rao and Molina, 2015).

Nonetheless, for each domain, these estimates are based only on data from the domain in question (curves and explanatory variables) without considering the rest of the sample. Like the Horvitz-Thompson estimator, they are therefore inaccurate for small domains and cannot be calculated for unsampled domains.

The methods that we present in the following section will allow us, by presenting a model common to all units of the population that describes the link between variables of interest and auxiliary information, to jointly use all data from the sample to perform the estimate for each domain, and thus increase the accuracy for each one. It will also make it possible to even provide estimates for unsampled domains.

4 Model-based estimation methods

In this section, we use the model from Valliant, Dorfman and Royall (2000), in which curves Y_i are considered to be random and we propose four innovative approaches for responding to our problem estimating curves with a mean demand for small domains. Assuming that $Y_i(t)$ and the auxiliary information vector $\mathbf{X}_i$ are available for each individual i in the domain d and that the mean $\overline{\mathbf{X}}_d = \sum_{i \in U_d} \mathbf{X}_i \big/ N_d$ is also known.

We assume that the auxiliary variables are related to the demand curves according to a functional model of superpopulation on all of the population that is generally expressed as:

$$\xi: \quad Y_i(t) = f_d(\mathbf{X}_i, t) + \epsilon_i(t), \quad i \in U_d, \quad t \in [0, T], \tag{4.1}$$

with f_d, an unknown regression function to be estimated, which can vary from one domain to another, and ϵ_i a process of zero expectation noise, zero covariance for different individuals and non-null regarding time.

If the size of domain N_d is large, then the mean μ_d will be estimated by

$$\hat{\mu}_d(t) = \frac{1}{N_d} \sum_{i \in U_d} \hat{Y}_i(t), \quad t \in [0, T],$$

where $\hat{Y}_i(t) = \hat{f}_d(\mathbf{X}_i, t)$ is the prediction of $Y_i(t)$. Otherwise, the mean μ_d is estimated by (see Valliant et al., 2000):

$$\hat{\mu}_d(t) = \frac{1}{N_d}\left(\sum_{i \in s_d} Y_i(t) + \sum_{i \in U_d - s_d} \hat{Y}_i(t)\right), \quad t \in [0, T]. \tag{4.2}$$

The quality of our estimates thus depends on the quality of our model: if the model is false, that may lead to biases in the estimates.

4.1 Functional linear model

The simplest model of form (4.1) is the functional linear regression model from Faraway (1997):

$$Y_i(t) = \mathbf{X}_i' \boldsymbol{\beta}(t) + \varepsilon_i(t), \quad t \in [0, T], \quad i \in U_d. \tag{4.3}$$

where the residuals $\varepsilon_i(t)$ are independent for $i \neq j$, distributed based on a law of means of 0 and of variance of $\sigma_i^2(t)$. If the size of domain N_d is large, then the mean of Y in domain d is estimated by

$$\hat{\mu}_d^{\text{blu}}(t) = \overline{\mathbf{X}}_d' \hat{\boldsymbol{\beta}}_{\text{BLU}}(t), \quad t \in [0, T],$$

where $\hat{\boldsymbol{\beta}}_{\text{BLU}}(t) = \left(\sum_{i \in s} \mathbf{X}_i \mathbf{X}_i' \big/ \sigma_i^2(t)\right)^{-1} \sum_{i \in s} \mathbf{X}_i Y_i(t) \big/ \sigma_i^2(t)$ is the best linear unbiased (BLU) estimator of $\boldsymbol{\beta}$ that does not depend on domain d. Estimator $\hat{\mu}_d^{\text{blu}}(t)$ can be expressed as a weighted sum of $Y_i(t)$:

$$\hat{\mu}_d^{\text{blu}}(t) = \frac{1}{N_d} \sum_{i \in s} w_{id}^{\text{blu}}(t) Y_i(t), \quad t \in [0, T],$$

where the weight $w_{id}^{\text{blu}}(t) = \left(\sum_{j \in U_d} \mathbf{X}_j'\right)\left(\sum_{j \in s} \mathbf{X}_j \mathbf{X}_j' \big/ \sigma_j^2(t)\right)^{-1} \mathbf{X}_i \big/ \sigma_i^2(t)$ now dependant on time t. If $n_d \big/ N_d$ is not insignificant, then the mean μ_d is estimated using (4.2) by:

$$\hat{\mu}_d^{\text{blu}}(t) = \frac{1}{N_d}\sum_{i\in s_d}\left(Y_i(t) - \mathbf{X}_i^T\hat{\boldsymbol{\beta}}_{\text{BLU}}(t)\right) + \overline{\mathbf{X}}_d'\hat{\boldsymbol{\beta}}_{\text{BLU}}(t), \qquad t\in[0, T].$$

This estimator can again be expressed as a weighted sum of $Y_i(t)$ with weights that will still depend on time t. The variance function (based on the model) of $\hat{\mu}_d^{\text{blu}}(t)$ can be derived using Rao and Molina (2015), Chapter 7. The variance function $\sigma_i^2(t)$ is unknown and can be estimated by following Rao and Molina (2015). By replacing $\sigma_i^2(t)$ with $\hat{\sigma}_i^2(t)$, we will obtain the empirical best linear unbiased predictor (EBLUP) of μ_d and its variance can be obtained using the method set out by Rao and Molina (2015). This EBLUP estimator does not use the sample weight d_i and therefore is not consistent in terms of the sampling plan (unless the sample weights are constant for units in the same domain d). A modified estimator, also referred to as a pseudo-EBLUP, can be constructed using the new approach described by Rao and Molina (2015, Chapter 7), equal in this case to the estimator set out in (3.4).

If $Y_i(t)$ is unknown for the units in domain d, the following indirect estimator can be used:

$$\hat{\mu}_d^{\text{ind}}(t) = \overline{\mathbf{X}}_d\hat{\boldsymbol{\beta}}(t) = \frac{1}{N_d}\sum_{i\in s}\tilde{w}_{id}^{\text{ind}}Y_i(t), \qquad t\in[0, T], \tag{4.4}$$

with $\hat{\boldsymbol{\beta}}(t)$ given in (3.5) and the weights $\tilde{w}_{id}^{\text{ind}} = \left(\sum_{j\in U_d}\mathbf{X}_j'\right)\left(\sum_{i\in s}d_i\mathbf{X}_i\mathbf{X}_i'\right)^{-1}\sum_{i\in s}d_i\mathbf{X}_i$ are not dependant on time t, unlike w_{id}^{blu}. Thus, the estimators proposed in this section have the benefit of being able to be used for unsampled domains.

4.2 Unit-level linear mixed models for functional data

The unit-level linear mixed models proposed by Battese, Harter and Fuller (1988) are very useful in estimating total actual variables for domains. As we will see later in more detail, they can translate both the effect of auxiliary information on the interest variable (by fixed effects), and the specifics of the domains (by random effects).

In this section, we thus attempt to adapt those models to the context of functional data. To that end, we will project curves in a space of defined dimensions and in that way, transform our functional problem into several problems in estimating total or mean real uncorrelated variables for small domains, which we will then resolve using the usual methods. The use of projection bases thus makes it possible to preserve the temporal correlation structure of our data while arriving at several unrelated subproblems in estimating real variables, which we treat independently using the usual methods.

4.2.1 Estimation of curves using unit-level linear mixed models applied to PCA scores

Like PCA in finite dimensions, functional PCA is a dimension-reduction method that makes it possible to summarize information contained in a data set. It was proposed by Deville (1974), its theoretical properties were studied in Dauxois, Pousse and Romain (1982) or Hall, Mülller and Wang (2006) and, finally, it was adapted to surveys by Cardot et al. (2010).

More formally, the curves $Y_i = (Y_i(t))_{t \in [0, T]}$ are functions of time t and we assume that they belong to $L^2[0, T]$, the space of square-integrable functions in the interval $[0, T]$. That space is equipped with the usual scalar product $< f, g >= \int_0^T f(t)\, g(t)\, dt$ and the standard $\| f \| = \left(\int_0^T f^2(t) dt \right)^{1/2}$. The variance covariance function $v(s, t)$ defined by:

$$v(s, t) = \frac{1}{N} \sum_{i=1}^{N} (Y_i(s) - \mu(s))(Y_i(t) - \mu(t)), \quad s, t \in [0, T], \tag{4.5}$$

with $\mu = \sum_{i=1}^{N} Y_i \big/ N$ the mean curve of Y on the population U.

Let $(\lambda_k)_{k=1}^{N}$ the eigen values of v with $\lambda_1 \geq \lambda_2 \ldots \geq \lambda_N \geq 0$ and $(\xi_k)_{k=1}^{N}$ the related orthonormal eigen vectors, $v(s, t) = \sum_{k=1}^{N} \lambda_k \xi_k(s)\, \xi_k(t)$.

The best approximation of Y in a dimensional space K smaller than N is given by the projection of Y in the space created by the first eigen vectors ξ_k, $k = 1, \ldots, q$ (Ramsay and Silverman, 2005):

$$Y_i(t) = \mu(t) + \sum_{k=1}^{K} f_{ik} \xi_k(t) + R_i(t), \quad i \in U, \quad t \in [0, T], \tag{4.6}$$

where f_{ik} is the projection (or score) of Y_i on the component ξ_k and $R_i(t)$, the rest representing the difference between curve i and its projection. The score f_{ik} is independent of the domain and can be calculated as the scalar product between ξ_k and $Y_i - \mu$, $f_{ik} =< Y_i - \mu, \xi_k >= \int_0^T (Y_i - \mu)(t)\, \xi_k(t)\, dt$. The decomposition given in (4.6) is also known as Karhunen-Loève.

Using (4.6), the mean μ_d on the domain d can be approximated by

$$\mu_d(t) \simeq \mu(t) + \sum_{k=1}^{K} \left(\frac{1}{N_d} \sum_{i \in U_d} f_{ik} \right) \xi_k(t), \quad d = 1, \ldots, D, \quad t \in [0, T]. \tag{4.7}$$

The unknown mean μ is estimated using the Horvitz-Thompson estimator

$$\hat{\mu}(t) = \frac{1}{N} \sum_{i \in s} d_i Y_i(t), \quad t \in [0, T] \tag{4.8}$$

and the ξ_k, $k = 1, \ldots, K$ are estimated by $\hat{\xi}_k$, the eigen vectors in the estimated variance-covariance function $\hat{v}(s, t) = \sum_{i \in s} d_i (Y_i(s) - \hat{\mu}(s))(Y_i(t) - \hat{\mu}(t)) \big/ N$ (Cardot et al., 2010).

Thus, to estimate μ_d, we must estimate the mean scores on the principal components for the domain d, i.e., $\overline{f}_{dk} = \sum_{i \in U_d} f_{ik} \big/ N_d$. To that end, for each component f_{ik}, $k = 1, \ldots, K$, we consider a unit-level linear mixed model, also known as a nested error regression model (Battese et al., 1988):

$$f_{ik} = \boldsymbol{\beta}_k' \mathbf{X}_i + \nu_{dk} + \epsilon_{ik}, \quad i \in U_d, \quad k = 1, \ldots, K, \tag{4.9}$$

with $\boldsymbol{\beta}_k' \mathbf{X}_i$ the fixed effect of auxiliary information, ν_{dk} the random effect of the domain d and ϵ_{ik} the residual of unit i. We assume that the random effects of the domains are independent and follow a common law of means of 0 and of variance of $\sigma_{\nu k}^2$. The residuals are also independent, distributed based on a law of means of 0 and of variance of $\sigma_{\epsilon k}^2$. In addition, the random effects and residuals are also assumed to be independent. The parameter $\boldsymbol{\beta}$ in the model can be estimated by $\tilde{\boldsymbol{\beta}}_k$, the best linear unbiased estimator

(BLUP) (Rao and Molina, 2015, Chapter 4.7) and the BLUP estimator $\overline{f}_{dk}$ is thus expressed as a composite estimator (see Rao and Molina, 2015):

$$\tilde{\overline{f}}_{dk} = \gamma_k \left(\overline{f}_{dk,s} - \left(\overline{\mathbf{X}}_{d,s} - \overline{\mathbf{X}}_d \right)' \tilde{\boldsymbol{\beta}}_k \right) + (1 - \gamma_k) \overline{\mathbf{X}}'_d \tilde{\boldsymbol{\beta}}_k, \quad k = 1, \ldots, K \tag{4.10}$$

with $\gamma_k = \sigma^2_{\nu k} / (\sigma^2_{\nu k} + \sigma^2_{\epsilon k} / n_d)$ and $\overline{\mathbf{X}}_{d,s} = \sum_{i \in s_d} \mathbf{X}_i / n_d$, $\overline{f}_{dk,s} = \sum_{i \in s_d} f_{ik} / n_d$ the respective means of the vectors $\mathbf{X}_i$ and the scores $\hat{f}_{ik}$ on s_d. Finally, the mean μ_d is estimated by

$$\hat{\mu}^{\text{BHF}}_d (t) = \hat{\mu}(t) + \sum_{k=1}^{K} \widehat{\overline{f}}_{dk} \hat{\xi}_k (t), \quad t \in [0, T], \tag{4.11}$$

with $\hat{\mu}$ and $\hat{\xi}_k$ the estimates of μ and k^{th} principal component ξ_k given previously.

The variances $\sigma^2_{\nu k}$ and $\sigma^2_{\epsilon k}$ for $k = 1, \ldots, K$ are unknown and are estimated by $\hat{\sigma}^2_{\nu k}$ and $\hat{\sigma}^2_{\epsilon k}$ obtained, for example, by restricted maximum likelihood (Rao and Molina, 2015). The estimator for $\overline{f}_{dk}$ is obtained by replacing γ_k in (4.10) with $\hat{\gamma}_k = \hat{\sigma}^2_{\nu k} / (\hat{\sigma}^2_{\nu k} + \hat{\sigma}^2_{\epsilon k} / n_d)$ and is known as empirical best linear unbiased prediction (EBLUP). Nonetheless, the calculation of the variance function (based on the model) of $\hat{\mu}^{\text{BHF}}_d (t)$ is more complicated in this case because of the estimators $\hat{\xi}_k$ of the principal components ξ_k and will be examined elsewhere.

We note that a simpler model, without the random effects, could have been considered for the scores f_{ik}:

$$f_{ik} = \boldsymbol{\beta}'_k \mathbf{X}_i + \epsilon_{ik}, \quad i \in U, \quad k = 1, \ldots, K, \tag{4.12}$$

with ϵ_{ik} a null mean residual σ^2_k. In this case, the parameter $\boldsymbol{\beta}_k$ is estimated by $\hat{\boldsymbol{\beta}}_k$, the BLUP estimator and the mean score on the domain d by $\widehat{\overline{f}}_{dk} = \hat{\boldsymbol{\beta}}'_k \overline{\mathbf{X}}_d$, $k = 1, \ldots, K$.

If the rate of n_d / N_d is not insignificant, then $\hat{\mu}_d$ is obtained using the procedure described in Section 4.1.

Note 1: Here, the PCA is not used as a dimension-reduction method, but to decompose our problem into several unrelated subproblems in the estimation of total real variables, which we know how to resolve. We thus keep a number K of principle components as high as possible, i.e., equal to the minimum number of instants of discretization and the number of individuals in the sample.

Note 2: When certain explanatory variables in vector $\mathbf{X}_i$ are categorical, our method, defined in the case of real variables, must be adapted: to that end, we propose transforming each categorical variable into a series of one hot encoding indicators $0 - 1$. As well, when the number of explanatory variables p is large, it may also be relevant to introduce penalties, RIDGE-style for example, in the regression problem.

Note 3: Other projection bases can be considered, such as wavelets (see Mallat, 1999), as they are particularly suited to irregular curves. Finally, another solution would be to apply the functional linear mixed models for curve values to the instants of discretization; however, that method would not allow for consideration of temporal correlations in the problem, unlike the previous ones.

4.2.2 Estimating variance by parametric bootstrap

To estimate the accuracy (variance based on the model) of mean curve estimators, we propose declining the parametric bootstrap method proposed by González-Manteiga, Lombarda, Molina, Morales and Santamara (2008) and then reiterated by Molina and Rao (2010). This is a replicate method that consists of generating a large number B of replicates $s^{*(b)}$, $b = 1, \ldots, B$ of size n by simple random sampling with replacement in s and randomly generating the random and fixed effects in the estimated superpopulation model. Note $\hat{R}_i(t) = Y_i(t) - \hat{\mu}(t) + \sum_{k=1}^{K} f_{k,i}\hat{\xi}_k(t)$ the estimated projection residual for the unit $i \in s$ (see also the formula in (4.6)). For $b = 1, \ldots, B,$ we proceed as follows for each $t \in [0, T]$:

1. Generate the random bootstrap effects of each domain, for each principal component:

$$\nu_{k,d}^{*(b)} \sim \mathcal{N}\left(0, \hat{\sigma}_{k,\nu}^2\right), \quad d = 1, \ldots, D, \quad k = 1, \ldots, K$$

   and, independent of those random effects, generate the individual bootstrap errors for each unit $i = 1, \ldots, n$ and for each principal component:

$$\epsilon_{k,i}^{*(b)} \sim \mathcal{N}\left(0, \hat{\sigma}_{k,\epsilon}^2\right), \quad i = 1, \ldots, n, \quad k = 1, \ldots, K.$$

2. Calculate the n bootstrap replicates of the projection residuals $\hat{R}_i^{*(b)}(t)$, for $i \in s^{*(b)}$ (this means selection with replacement of n projection residuals among the n residuals $\hat{R}_i(t)$).
3. Calculate the bootstrap replicates $Y_i^{*(b)}(t)$ conditional to the explanatory variables $\mathbf{X}_i$ using the estimated model:

$$Y_i^{*(b)}(t) = \hat{\mu}(t) + \sum_{k=1}^{K} \underbrace{\left(\mathbf{X}_i'\hat{\beta}_k + \nu_{k,d}^{*(b)} + \epsilon_{k,i}^{*(b)}\right)}_{f_{k,i}^{*(b)}} \hat{\xi}_k(t) + \hat{R}_i^{*(b)}(t), \quad \forall i \in s_d^{*(b)} = s^{*(b)} \cap s_d.$$

   We see that $f_{k,i}^{*(b)}$, the simulated score for the unit $i,$ is obtained using the same approach as in González-Manteiga et al. (2008).
4. For each domain d, calculate the bootstrap replicate $\hat{\mu}_d^{*(b)}$ on the replicate $s^{*(b)}$ by declining the entire process: PCA and estimation of linear mixed models on principal components by means of EBLUP.
5. Estimate the estimator's variance $\hat{\mu}_d(t)$ by the empirical variance of the B replicates $\hat{\mu}_d^{*(b)}$:

$$\hat{V}(\hat{\mu}_d(t)) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\mu}_d^{*(b)}(t) - \frac{1}{B}\sum_{b=1}^{B}\hat{\mu}_d^{*(b)}(t)\right)^2.$$

This approach will also be the one used to approximate the variance of the functional linear regression (omitting step 1 of generating random effects $\nu_{k,d}^{*}$).

4.3 Non-parametric estimation using regression trees and random forests for small curve domains

In this section, to obtain individual predictions $\hat{Y}_i(t)$, we use non-parametric models, regression trees adapted to functional data, and random forests, which no longer require a linear form in the relation between auxiliary information and interest variable and allow more flexibility in modelling. In fact, regression trees for functional data are frequently used at EDF and are known to give satisfactory results on electricity consumption curves. As well, in literature, regression trees have been adapted to surveys by Toth and Eltinge (2011), but not for estimating totals in small domains.

In this section and the next section, we are therefore seeking to estimate a specific case of the general model (4.1) in which the function f does not depend on the domain of the unit i,

$$Y_i(t) = f(X_i, t) + \epsilon_i(t) \quad \forall i \in U, \quad t \in [0, T]. \tag{4.13}$$

4.3.1 Regression trees for functional data

The classification and regression tree (CART) proposed by Breiman, Friedman, Stone and Olshen (1984) is a very popular non-parametric statistical technique. Its goal is to predict the value of a real or categorical target variable Y based on a vector of real or categorical explanatory variables $\mathbf{X} = (X_1, \ldots, X_j, \ldots, X_p)$. To that end, we determine a partitioning of the space of $\mathbf{X}$ by repeatedly splitting the data set in two, using the decision rule (split criteria) involving a single explanatory variable. Thus, our sample s is the first node λ in a tree (its "root") that we seek to subdivide into two separate nodes λ_l and λ_r such that the values of the real target variable Y_i are as homogenous as possible in each node. The inertia criterion $\kappa(\lambda)$ used to quantify the homogeneity of a node is frequently the sum of the squares of residuals between the values of Y_i for units i in node λ and the mean of those values in the node: $\kappa(\lambda) = \sum_{i \in \lambda} \left(Y_i - \overline{Y}_\lambda\right)^2$ where $\overline{Y}_\lambda$ is the mean of Y_i in node λ.

For the variables X_j which are quantitative, the decision rules are expressed as

$$\begin{cases} i \in \lambda_l & \text{if} \quad X_{ji} < c \\ i \in \lambda_r & \text{otherwise,} \end{cases} \tag{4.14}$$

with c a cut-off point to be optimized among all possible values of X_j. For qualitative variables, they consist of dividing into two separate subsets of modalities. The search for the optimal split criterion is a matter of resolving the optimization problem

$$\underset{\lambda_l, \lambda_r}{\arg\max} \, (\kappa(\lambda) - \kappa(\lambda_l) - \kappa(\lambda_r)). \tag{4.15}$$

Each of these nodes will then be subdivided in turn into two child nodes and the partitioning process continues until a minimal node size is obtained, until the value of the target variable is the same for all units of the node, or until a given maximum depth is attained. The final partition of the space is then made up of the final nodes in the tree, also called leaves. A summary of each of those leaves (often the mean for a quantitative target variable) then becomes the predicted variable for all units assigned to the leaf. The various parameters (minimum node size and depth) can be selected by cross-validation.

When the variable Y to be predicted is not a real variable, but a dimension vector $m > 1$, the regression tree principle extends very naturally: the tree construction algorithm and the choice of cross-validation parameters remain unchanged, but the inertia criterion is modified. Thus, the minimization problem is still in the form of (4.15), but this time the criterion is in the form of $\kappa(\lambda) = \sum_{i\in\lambda} \left\| Y_i - \overline{Y}_\lambda \right\|^2$ where $\|\cdot\|$ is, for example, the Euclidean norm or the Mahalanobis distance norm. Multivariate regression trees were used, for example, by De'Ath (2002) in an ecological application.

Finally, when the variable to be predicted Y is a curve, the algorithm for construction of the tree and for choosing the parameters is the same, but this time a functional inertia criterion κ must be used. There are many possible choices. We chose to follow the "Courbotree" approach described in Stéphan and Cogordan (2009) and frequently used at EDF for building segmentations of data sets of electricity consumption curves based on explanatory variables. In that approach, we apply the method presented in the previous paragraph for multivariate Y on vectors $\mathbf{Y}_i = (Y_i(t_1), \ldots, Y_i(t_L))$ of curve values at the instants of discretization, with the Euclidian distance. The Euclidian distance on instants of discretization can thus be seen as an approximation of the norm $L^2[0, T]$. More formally, the functional criterion is thus expressed as

$$\kappa(\lambda) = \sum_{i\in\lambda}\sum_{l=1}^{L}\left(Y_i(t_l) - \overline{Y}_\lambda(t_l)\right)^2, \tag{4.16}$$

with $\overline{Y}_\lambda(t_l) = \sum_{i\in\lambda} Y_i(t_l) \big/ n_\lambda$ where n_λ is the number of units in the sample that belong to the node λ.

In practice, when working on electricity consumption data, the curves considered are at extremely similar levels, and the Courbotree algorithm based on the Euclidian distance may not work well when applied to raw data. Often, the Courbotree algorithm is therefore only used on the curve forms, i.e., the normalized curves $\tilde{Y}_i(t) = Y_i(t) \big/ \overline{Y}_i$ where $\overline{Y}_i = \sum_{\ell=1}^{L} Y_i(t_\ell) \big/ L$ is the mean of $Y_i(t)$ (or the level) on all measurement instants $t_1, \ldots, t_L$ (method also known as *normalized Courbotree*). We then calculate the prediction $\overline{Y}_i$ using a linear regression and finally obtain the prediction of $Y_i(t)$ by obtaining the product between the prediction of $\tilde{Y}_i(t)$ and that of $\overline{Y}_i$.

4.3.2 Variance estimation

To estimate the variance under the model of our estimators for mean curves by domain, we will follow a bootstrap approach very similar to the one proposed for parametric models. Here, our superpopulation model is expressed as

$$Y_i(t) = f(\mathbf{X}_i, t) + \epsilon_i(t), \quad \forall i \in U. \tag{4.17}$$

Let $\hat{f}(\mathbf{X}_i, t)$, for all $i \in s$ the predicted value for the unit i by regression tree, and $\hat{\epsilon}_i(t) = Y_i(t) - \hat{f}(\mathbf{X}_i, t)$, for all $i \in s$ the estimated residual for that unit. The idea of our accuracy approximation method is, as with linear mixed models, to generate a large number B of replicates $s^{*(b)}$, $b = 1, \ldots, B$ of size n by simple random sampling with replacement in s, and calculate the estimator of the mean curve by domain on each replicate and, finally, deduct the variance from the estimator by the variability between replicates. The bootstrap method used here is also known as residual bootstrap in linear model cases.

More specifically, for $b = 1, \ldots, B$, we proceed as follows for each $t \in [0, T]$:

1. Calculate the bootstrap replications of the estimated residuals $\hat{\epsilon}_i^{*(b)}(t)$ for $i \in s^{*(b)}$.
2. Calculate the bootstrap replications for $Y_i(t)$:

$$Y_i^{*(b)}(t) = \hat{f}(\mathbf{X}_i, t) + \hat{\epsilon}_i^{*(b)}(t), \qquad \forall i \in s^{*(b)}$$

and recalculate, for each domain d, the mean estimator $\hat{\mu}_d^{*(b)}(t)$ on this replicate.

3. Estimate the variance by the empirical variance of the B bootstrap replicates $\hat{\mu}_d^{*(b)}(t)$, $\hat{V}(\hat{\mu}_d(t)) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\mu}_d^{*(b)}(t) - \frac{1}{B}\sum_{b=1}^{B}\hat{\mu}_d^{*(b)}(t)\right)^2.$

The process is identical if we estimate the function f by random forests rather than regression trees.

4.4 Aggregation of predictions by random forests for curves

The literature often highlights the mediocre predictive performances of regression trees compared to other techniques such as SVMs (see, for example, Cristianini and Shawe-Taylor, 2000). Regression trees can be unstable and highly dependant on the sample on which they were built. To resolve that default, Breiman (2001) proposed the random forest algorithm. This is a set technique that, as its name suggests, consists of aggregating predictions resulting from different regression trees. The fact that the aggregation of unstable predictors leads to a reduction in variance was particularly shown by Breiman (1998). For a quantitative target variable, the aggregation of predictions is performed by taking the mean predictions for each of the trees.

To reduce the variance of the aggregate prediction, the objective is to build trees that are very different from each other. The Breiman algorithm introduces variability in the construction of the trees on the one hand by means of replication (simple random sampling with replacement) of units and, on the other hand, by means of random selection, for each "split" in the tree, of a subset of candidate explanatory variables. For a regression tree, there are therefore two additional parameters to be adjusted for a random forest: the number of trees and the number of candidate explanatory variables in each split.

When the interest variable is multivariate (or functional), the algorithm proposed by Breiman adapts easily, by aggregating the multivariate (or functional) regression trees presented in the previous paragraph. Multivariate random forests have, for example, been studied by Segal and Xiao (2011).

The algorithm that we are proposing here, called "Courboforest," simply consists of aggregating the functional regression trees constructed using the "Courbotree" approach, i.e., the multivariate regression trees applied to the vectors $(Y_i = Y_i(t_1), \ldots, Y_i(t_L))$ of the values of the curves at the instants of discretization, with the split criterion being the inertia based on the Euclidean distance defined by equation (4.16).

5 Application to electricity consumption curves

We will now test the methods that we have just presented to compare their performance on electricity consumption data for French residential clients.

5.1 Presentation of the data set

We worked with a data set belonging to EDF that contains electricity consumption curves for $N = 1{,}905$ French residential clients by daily interval from October 2011 to March 2012, without any missing values $(L = 177$ points). This population is subdivided into $D = 8$ domains corresponding to geographic areas with respective sizes of 573, 195, 304, 121, 228, 219, 45 and 220. For confidentiality purposes, we cannot describe the data set in great detail, or show the mean curves by domain.

By way of illustration, Figure 5.1 shows the appearance of the standardized curves (i.e., each curve is divided by its mean calculated over the period of time studied) for five random individuals, and Figure 5.2 shows the appearance of the first five principal components of the functional PCA created for this data set.

We see that the first component, the overall appearance of which is similar to that of the mean curve, is a "level" effect. Components two and three, which present peaks during the coldest period in February, describe the sensitivity of consumption to outside temperatures. The fourth compares "mid-season" consumption to "wintertime" consumption and, finally, the fifth shows a low at about Christmas (and about February 14).

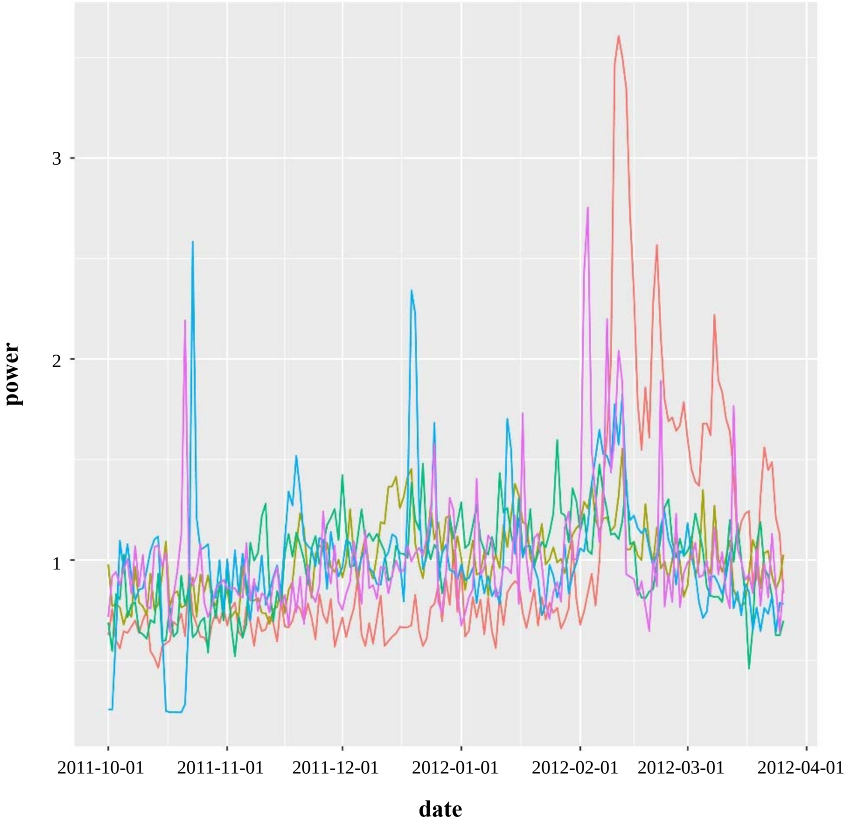


Figure 5.1 Standardized electricity consumption curves (i.e., divided by their mean over the study period) by daily interval for residential clients, winter, 2011/2012.

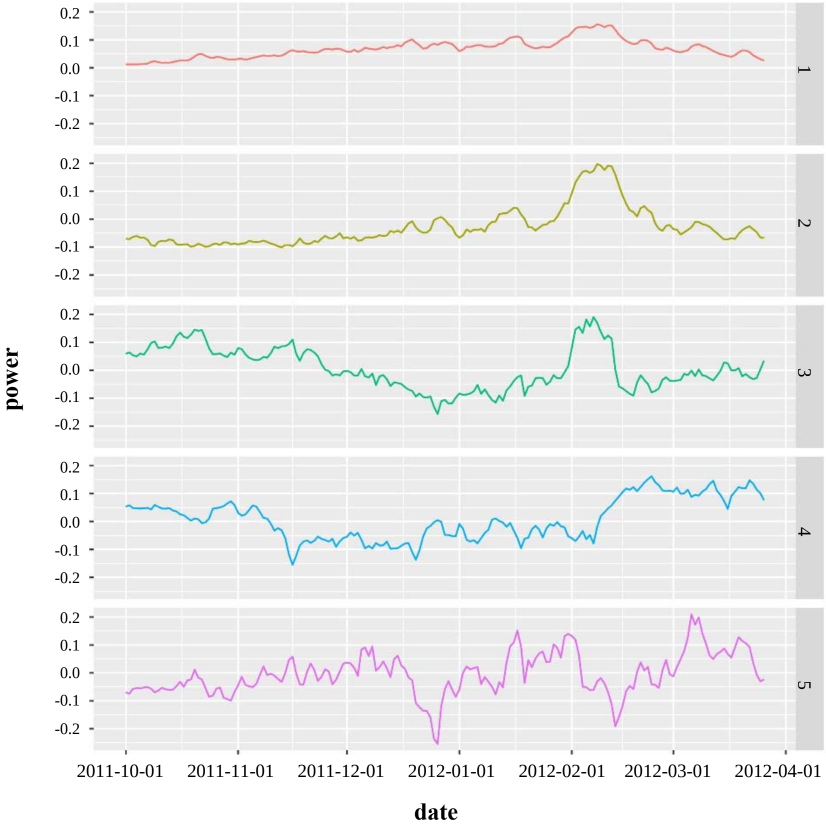


Figure 5.2 First five components of the principal component analysis.

For each individual in our population of study, we have four auxiliary variables at the individual level: contract power (in three classes), rate option (base or off-peak periods) (in the base option, the price per kWh remains constant, while the rate for off-peak periods is reduced for eight hours [referred to as off-peak]. The largest consumers tend to prefer that rate. Off-peak periods can vary from one client to another, but this factor has no impact here, as we are working on a daily interval), the previous year's annual consumption, and the type of dwelling (apartment or single home). These auxiliary variables remain the same for all methods used in order to compare identical auxiliary information. All tests were implemented in R.

5.2 Test protocol

We compare various estimators obtained using the methods set out in this chapter, for various types of modelling (unit-level linear mixed models, linear functional regressions, regression trees, random forests). We test two versions of the unit-level linear mixed model, one by placing linear mixed models on the PCA scores, as suggested in Section 4.2, and the other by applying them directly to the values of the curves of instants of discretization.

For non-parametric methods, the forests and trees have a depth (number of levels) of 5 and a minimum size of 5 leaves. There are 40 trees in the forests. The algorithms can be applied by separating the estimation of the level of the curve and its form (standardization = "yes") or not separating (standardization = "no"). To not multiply the possible combinations, we finally focused on the estimators listed in Table 5.1. The parameters of the regression tree and random forest models are set out in Table 5.2.

Table 5.1
Various estimation method tests

Title	Reference	Projection
Horvitz-Thompson	Equation (3.1)	None
Calibration	Equation (3.2)	None
Linear mixed model	Section (4.2)	None
Linear mixed model on PCA	Equation (4.11)	PCA
Linear regression	Equation (4.4)	None
Courbotree	Section (4.3)	None
Standardized Courbotree	Section (4.3)	None
Courboforest	Section (4.4)	None

Table 5.2
Parameters for trees and random forests

Title	Depth (number of levels)	Number of trees	Standardization
Courbotree	5	1	No
Standardized Courbotree	5	1	Yes
Courboforest	5	40	No

To evaluate the quality of our estimation methods, our test protocol consists of conducting a large number E of sampling simulations from our original population and then estimating the mean curve for each $D = 8$ domain based on each sample gathered by the various proposed methods. In our simulations, the eighth domain $(d = D = 8)$ will always be unsampled in order to measure the performance of our various estimators in this scenario. For each simulation, we select $n = 200$ individuals by simple random sampling from among those in the seven sampled domains $(d = 1, \ldots, 7)$.

Let $\mu_d(t_l)$ the mean curve for the domain d at the instant t_l and $\hat{\mu}_d(t_l)$ its estimator by a given method. We calculate the relative bias of $\hat{\mu}_d(t_l)$:

$$\mathrm{RB}(\hat{\mu}_d(t_l)) = 100 \frac{E_{\mathrm{MC}}[\hat{\mu}_d(t_l)] - \mu_d(t_l)}{\mu_d(t_l)}, \quad d = 1, \ldots, D, \quad l = 1, \ldots, L, \tag{5.1}$$

where $E_{\mathrm{MC}}[\hat{\mu}_d(t_l)] = \sum_{e=1}^{E} \hat{\mu}_d^{(e)}(t_l) \Big/ E$ is the Monte Carlo expectation of the estimator $\hat{\mu}_d(t_l)$ with $\hat{\mu}_d^{(e)}(t_l)$ the estimator of the mean curve obtained for the e^{th} simulation, for $e = 1, \ldots, E$. A second indicator, known as relative efficiency (RE), is calculated as follows:

$$\mathrm{RE}(\hat{\mu}_d)(t_l) = 100 \frac{\mathrm{MSE}_{\mathrm{MC}}(\hat{\mu}_d)(t_l)}{\mathrm{MSE}_{\mathrm{MC}}(\hat{\mu}_d^{\mathrm{HT}})(t_l)}, \quad d = 1, \ldots, D-1, \quad l = 1, \ldots, L. \tag{5.2}$$

where $\text{MSE}_{\text{MC}}\left(\hat{\mu}_d\left(t_l\right)\right) = \sum_{e=1}^{E}\left(\hat{\mu}_d^{(e)}\left(t_l\right) - \mu_d\left(t_l\right)\right)^2 \Big/ E$ is the Monte Carlo mean square error, $d = 1, \ldots, D,\ l = 1, \ldots, L.$ The lower the RE indicator, the more the estimator will be considered effective. An RE of 100 corresponds to an indicator as effective as the reference estimator.

Here, the reference estimator $\hat{\mu}_d^{\text{HT}}$ is the Horvitz-Thompson estimator (which, for our simple random sampling plan, is the simple mean of the curves in the domain considered); it corresponds to the model described by equation (3.1). This estimator cannot be calculated for the unsampled domain. The RE estimator is then obtained by dividing the MSE of the various estimators by the mean MSE of the Horvitz-Thompson estimator over the seven sampled domains, i.e.

$$\text{RE}\left(\hat{\mu}_D\right)\left(t_l\right) = 100\,\frac{\text{MSE}_{\text{MC}}\left(\hat{\mu}_D\right)\left(t_l\right)}{\overline{\text{MSE}}_{\text{MC}}^{\text{HT}}\left(t_l\right)}, \quad l = 1, \ldots, L, \tag{5.3}$$

with $\overline{\text{MSE}}_{\text{MC}}^{\text{HT}}\left(t_l\right) = \sum_{d=1}^{D-1} \text{MSE}_{\text{MC}}\left(\hat{\mu}_d^{\text{HT}}\right)\left(t_l\right),\ l = 1, \ldots, L.$

For each indicator and each instant t_l, the results obtained for the various sampled domains are then aggregated for all domains, $\text{RB}_{\text{ech}}\left(\hat{\mu}\right)\left(t_l\right) = \frac{1}{D-1}\sum_{d=1}^{D-1}\text{RB}\left(\hat{\mu}_d\right)\left(t_l\right)$ and $\text{RE}_{\text{ech}}\left(\hat{\mu}\right)\left(t_l\right) = \frac{1}{D-1}\sum_{d=1}^{D-1}\text{RE}\left(\hat{\mu}_d\right)\left(t_l\right)$ for $l = 1, \ldots, L,$ while the indicators obtained for the unsampled domain are used as-is.

Finally, to evaluate overall performance, we consider the mean of those indicators for all instants in the test period, while still separating the sampled domains from the unsampled domain. We also look at the calculation times of the various estimators.

5.3 Results and test conclusions

The test results of the methods are presented in Table 5.3 and illustrated in Figures 5.3 to 5.5.

Table 5.3
Mean method performance indicators (RB, RE) for all instants of discretization and domains, separating the unsampled domain from the others

Domain type	Method	RE (%)	RB (%)
Sampled	Horvitz-Thompson	100,00	0,25
Sampled	Calibration	37,13	-0,47
Sampled	Linear mixed model	*14,69*	0,60
Sampled	Linear mixed model PCA	15,40	0,67
Sampled	Linear regression	24,87	1,20
Sampled	Courbotree	20,54	0,80
Sampled	Standardized Courbotree	22,35	1,45
Sampled	Courboforest	24,66	0,62
Unsampled	Horvitz-Thompson		
Unsampled	Calibration		
Unsampled	Linear mixed model	*13,43*	4,66
Unsampled	Linear mixed model PCA	13,49	4,77
Unsampled	Linear regression	14,38	5,09
Unsampled	Courbotree	14,29	3,48
Unsampled	Standardized Courbotree	16,63	5,88
Unsampled	Courboforest	15,97	0,37

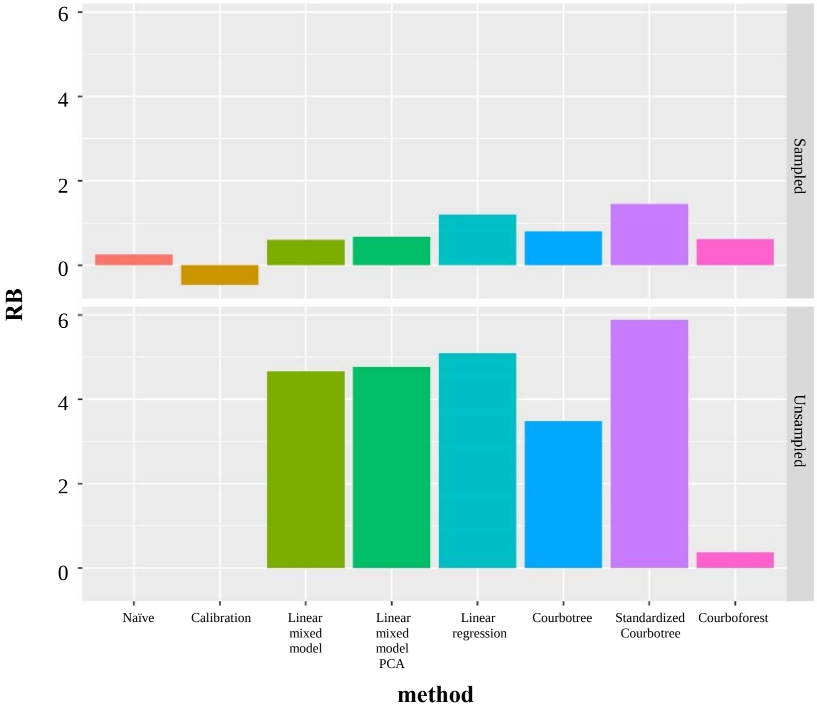


Figure 5.3 Mean relative biases as % (formula (5.1)) of estimation methods, for all instants in the domains, separating unsampled and sampled domains.

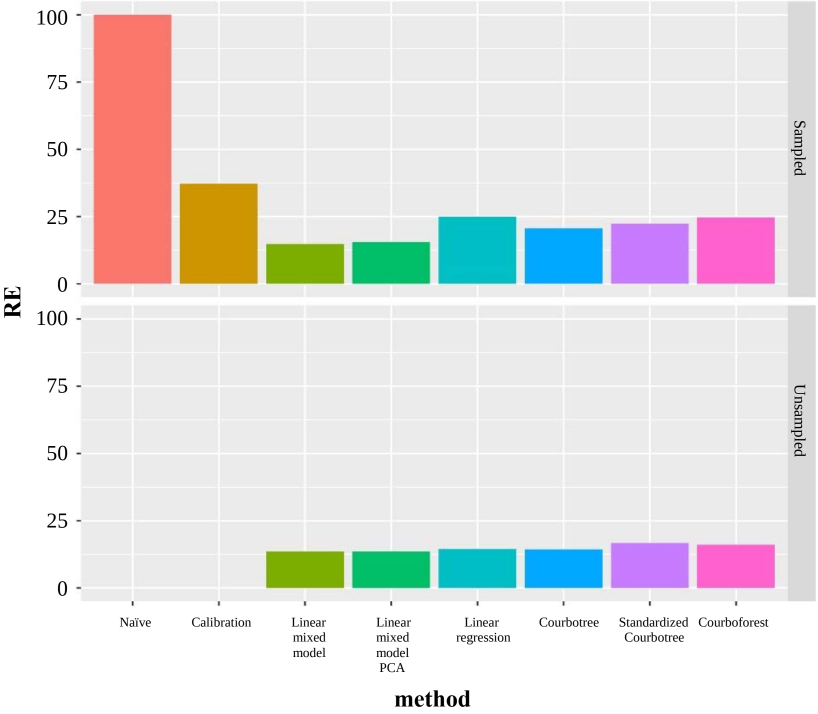


Figure 5.4 Mean relative efficiency (RE) (formula (5.2)) of the estimation methods for all instants and domains, separating unsampled and sampled domains.

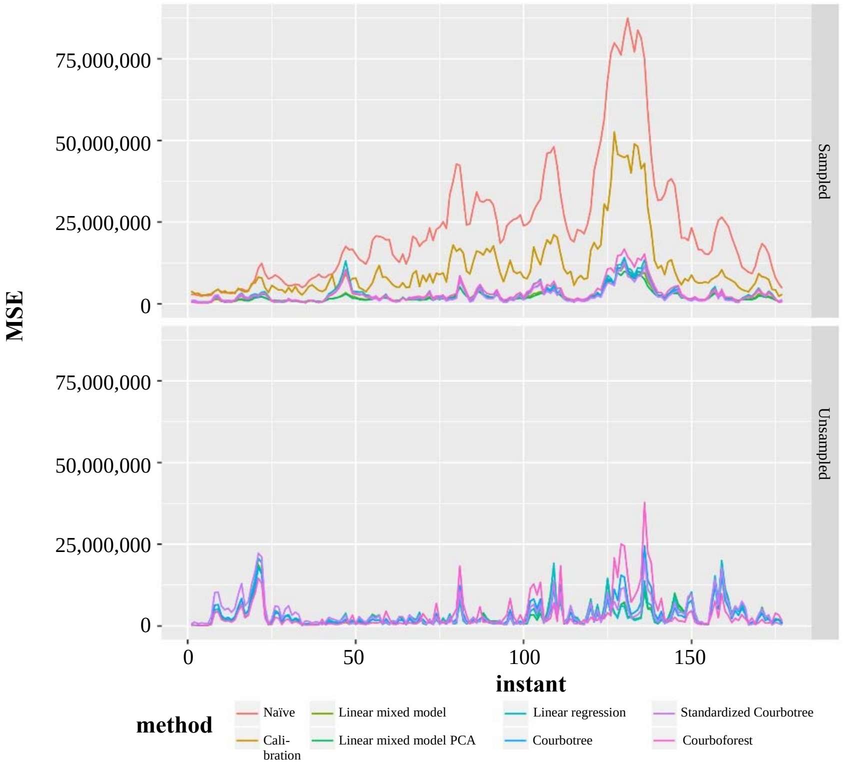


Figure 5.5 Evolution of the mean MSEs for domains over time, for the various estimation methods.

For sampled domains, we see that the integration of explanatory variables in the estimate, regardless of the method used, leads to a net gain in performance: thus, for the least effective method (the estimator by calibration), the error is divided by three when explanatory variables are used.

As well, the use of our various estimators based on superpopulation models leads to an additional gain in accuracy: the RE for our various methods thus range from 15% for linear mixed models to 25% for random forests.

The linear mixed models are the most effective method, so we can assume that there are characteristics of the domains that are unexplainable using only the auxiliary variables that this type of model is able to capture. We therefore go from an RE of 25% for the linear functional regression to an RE of approximately 15% by including these random effects.

The tree and random forest methods capture non-linearities in the relationship between explanatory variables and the interest variable, which explains why these methods give better results than linear functional regressions: the RE of the various non-parametric methods are between 20% and 25%, compared to 25% for linear functional regressions. Very surprisingly, the regression tree gives better results than the random forest. We can put forth the theory that this is because our objective is to best estimate the mean curve of a series of units, not each curve individually. It is therefore possible that the tree is not as good for predicting each curve, but better at the aggregate level. As well, on this particular data set, the method gives the best results when working on raw curves, not when distinguishing between the estimation of form and level.

Projecting curves based on the PCA does not seem to lead to any significant gains in accuracy here.

The Horvitz-Thompson estimator cannot be produced on unsampled domains. The differences between the other methods are much more restricted than on the sampled domains: the random effects cannot be estimated for unsampled domains.

Finally, in Figure 5.5, we trace the mean square error of our estimators for the sampled and unsampled domains. We note that this square error is higher in the winter (January and February). This high variability could be due to a sharp drop in outside temperatures during those months, which increases the variability of heating consumption (difference in behaviour and electrical heating equipment depending on clients). The naive and calibration estimators adapt least well to this situation.

5.4 Comparison of methods and selection criteria

Each model-based method has benefits and drawbacks. Unit-level linear mixed models are the only ones that, due to random effects, make it possible for the modelling to include domain characteristics not reflected in auxiliary information. It thus seems relevant to use them when assuming that the explanatory variables do not make it possible to explain all differences between domains.

The linear functional regression ignores the random effect of the domains, so we expect it to be less effective than linear mixed models due to its construction. Finally, the two non-parametric methods allow for better modelling of the non-linear relationships between the explanatory variables and the interest variable, but on the other hand, does not make it possible to capture the differences between domains that are not reflected in the auxiliary information. They also require the availability of auxiliary information $\mathbf{X}_i$ for each individual in the population when, in the past, we only needed mean values $\overline{\mathbf{X}}_d$ for each domain in the population and $\mathbf{X}_i$ for the sample. The choice between a parametric and non-parametric approach will therefore depend on the nature of the problem, the diversity of domains and the explanatory variables available. Be believe that neither of the two approaches is systematically preferable over the other.

A process for choosing between the two approaches could be to estimate the respective variances in the random effects and the residuals in the linear mixed models and, depending on the relative scope of those effects, moving more toward one or the other type of model. Conversely, cross-validation can be used to quantify the respective performance of the linear mixed models and the non-parametric models for predicting the aggregates of individual curves in order to direct our choice.

Among the non-parametric methods, the choice between regression trees and random forests will depend on the predictive performance of those methods on data, for the mean curves of domains. Generally, we can assume that random forests will give better results than regression trees for individual data (see Breiman et al., 1984); however, it is entirely possible that the best of the two methods for predicting each curve may not be the one that gives the best results to all domains or, at the very least, that the two methods are reduced when we consider the prediction of mean curves of individual aggregates. As well, due to their construction, random forests require a lot more calculation time than regression trees and that aspect cannot be ignored when the data sets being processed are large in size.

6 Conclusions and outlooks

In this article, we proposed four approaches for estimating mean curves by sampling for small domains. The first two consist of projecting curves in a finite space and using the usual methods for estimating total real variables for each base vector in the projection space. In this case, we use either unit-level linear mixed models or linear regression. The last two approaches consist of predicting each curve of the unsampled units using a non-parametric model and aggregating those predictions to determine the estimated mean curves for each domain. The models used to build the predictions are regression trees adapted to functional data build using the Courbotree approach of Stéphan and Cogordan (2009) or random forests adapted to functional data built by aggregating random Courbotree trees. For each approach, we also proposed a process for approximating the variance of mean curve estimators based on a bootstrap.

Our tests showed that the linear mixed models gave the best results and, for this particular data set, made it possible to divide the error committed by approximately seven in relation to the Horvitz-Thompson estimators. The regression trees come next, followed by the linear functional regressions.

This work can be extended in various ways. In particular, we feel that the approach based on the aggregation of non-parametric estimates of curves using regression trees or random forests is promising. An interesting possibility for improvement could be the use of more relevant distances than the Euclidean distance in the split criteria that builds our regression trees. We could thus use the Mahalanobis distance, the Manhattan distance, or a “dynamic time warping” distance.

Another possibility could be to build this split criterion by applying the Euclidian distance not on the discretized curves, but on a transformation of those curves, by projection in a wavelet base, or on non-linear summaries, such as variational autoencoders from deep learning models (see, for example, LeCun, Bengio and Hinton, 2015).

We can also question the choice of depth of the regression tree, the minimum size of the leaves and the number of trees in the forest. The criteria usually used in non-parametric statistics to answer this question are usually based on the principle of cross-validation. However, our objective here is not to determine the best possible prediction for each population unit, but a prediction that gives the best estimate of the mean curve by domain, which is not necessarily the same thing. It would therefore be best to adapt the cross-validation criteria to reflect our objective.

Finally, we note that the introduction of random effects in the linear models results in improved prediction, which leads us to think that there are characteristics in the domains that are not explained solely by the auxiliary information. It could therefore be relevant to adapt the functional regression trees to include the random effects. One solution, for example, would be to extend the algorithm from Hajjem, Bellavance and Larocque (2014), based on an EM algorithm as part of the functional data.

Acknowledgements

The authors thank Hervé Cardot for the fruitful discussions and the associate editor and two referees for their remarks and comments, which helped greatly improve this article.

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.

Breiman, L. (1998). Arcing classifiers (with a discussion and a response from the author). *The Annals of Statistics*, 26(3), 801-849.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.

Breiman, L., Friedman, J., Stone, C.J. and Olshen, R.A. (1984). *Classification and Regression Trees*. CRC press.

Cardot, H., Degras, D. and Josserand, E. (2013). Confidence bands for Horvitz-Thompson estimators using sampled noisy functional data. *Bernoulli*, 19(5A), 2067-2097.

Cardot, H., Goga, C. and Lardin, P. (2013). Uniform convergence and asymptotic confidence bands for model-assisted estimators of the mean of sampled functional data. *Electronic Journal of Statistics*, 7, 562-596.

Cardot, H., Chaouch, M., Goga, C. and Labruère, C. (2010). Properties of design-based functional principal components analysis. *Journal of Statistical Planning and Inference*, 140(1), 75-91.

Cardot, H., Dessertaine, A., Goga, C., Josserand, E. and Lardin, P. (2013). Comparison of different sample designs and construction of confidence bands to estimate the mean of functional data: An illustration on electricity consumption. *Survey Methodology*, 39, 2, 283-301. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2013002/article/11888-eng.pdf.

Cristianini, N., and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines*. Cambridge University Press Cambridge.

Dauxois, J., Pousse, A. and Romain, Y. (1982). Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference. *Journal of Multivariate Analysis*, 12(1), 136-154.

De’Ath, G. (2002). Multivariate regression trees: A new technique for modeling species-environment relationships. *Ecology*, 83(4), 1105-1117.

Deville, J.-C. (1974). Méthodes statistiques et numériques de l'analyse harmonique. In *Annales de l'INSEE*, JSTOR, 3-101.

Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American statistical Association*, 87(418), 376-382.

Faraway, J.J. (1997). Regression analysis for a functional response. *Technometrics*, 39(3), 254-261.

González-Manteiga, W., Lombarda, M.J., Molina, I., Morales, D. and Santamara, L. (2008). Analytic and bootstrap approximations of prediction errors under a multivariate Fay-Herriot model. *Computational Statistics & Data Analysis*, 52(12), 5242-5252.

Hajjem, A., Bellavance, F. and Larocque, D. (2014). Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation*, 84(6), 1313-1328.

Hall, P., Müller, H.-G. and Wang, J.-L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *The Annals of Statistics*, 1493-1517.

Horvitz, D.G., and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260), 663-685.

LeCun, Y., Bengio, Y. and Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444.

Mallat, S. (1999). *A Wavelet Tour of Signal Processing*. Academic press.

Molina, I., and Rao, J.N.K. (2010). Small area estimation of poverty indicators. *Canadian Journal of Statistics*, 38(3), 369-385.

Pfeffermann, D., and Burck, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology*, 16, 2, 217-237. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/1990002/article/14534-eng.pdf.

Ramsay, J.-O., and Silverman, B.-W. (2005). *Functional Data Analysis*. Springer Series in Statistics, New York, Second Edition.

Rao, J.N.K., and Molina, I. (2015). *Small Area Estimation*. New York: John Wiley & Sons, Inc.

Rao, J.N.K., and Yu, M. (1994). Small-area estimation by combining time-series and cross-sectional data. *Canadian Journal of Statistics*, 22(4), 511-528.

Segal, M., and Xiao, Y. (2011). Multivariate random forests. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1), 80-87.

Stéphan, V., and Cogordan, F. (2009). CourboTree: Application des arbres de régression multivariés pour la classification de courbes. *La Revue MODULAD*, June.

Toth, D., and Eltinge, J.L. (2011). Building consistent regression trees from complex sample data. *Journal of the American Statistical Association*, 106(496), 1626-1636.

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

Vol. 44, No. 2, pp. 215-238
Statistics Canada, Catalogue No. 12-001-X

Coordination of spatially balanced samples

Anton Grafström and Alina Matei[1]

Abstract

Sample coordination seeks to create a probabilistic dependence between the selection of two or more samples drawn from the same population or from overlapping populations. Positive coordination increases the expected sample overlap, while negative coordination decreases it. There are numerous applications for sample coordination with varying objectives. A spatially balanced sample is a sample that is well-spread in some space. Forcing a spread within the selected samples is a general and very efficient variance reduction technique for the Horvitz-Thompson estimator. The local pivotal method and the spatially correlated Poisson sampling are two general schemes for achieving well-spread samples. We aim to introduce coordination for these sampling methods based on the concept of permanent random numbers. The goal is to coordinate such samples while preserving spatial balance. The proposed methods are motivated by examples from forestry, environmental studies, and official statistics.

Key Words: Coordination; Local pivotal method; Spatially correlated Poisson sampling; Permanent random numbers; Unequal probability sampling designs; Transformed spatially correlated Poisson sampling.

1 Introduction

In the classical survey sampling framework, a random sample is selected from a finite population with a probability provided by the sampling design. The sampling design can be extended to the case of several samples, defining a joint probability to select them. On the other hand, two or more samples can be drawn from the same population or from overlapping populations, independently or not. Sample coordination applies to the latter case and seeks to create a probabilistic dependence between samples' selections based on a joint sampling design. It is used in the case of repeated surveys or of several surveys. Two types of coordination are defined in the literature: positive and negative. In the former case, the goal is to maximize the overlap between different samples. In the latter, one wants to minimize it. Positive coordination can be used to reduce the survey costs or to induce a positive covariance between successive estimators of state in repeated surveys, and thus reduce the variance of an estimator of change. Negative coordination may be applied to reduce the response burden of units that have a risk of being selected for several surveys.

When updating a sample in repeated surveys over time (a panel), deaths, births or merge of the units can appear in the population. Thus, the population changes over time and the same sample can not be used at each time occasion. New samples are drawn at different time occasions, but a certain degree of overlap between samples can be required. This can be achieved using positive coordination. On the other hand, negative coordination is usually used to draw samples in several surveys, involving thus different but overlapping populations. Due to births, deaths, changes in activity or size, splits, mergers, etc. of units in the same population or due to the use of different overlapping populations, an important problem in sample coordination is the difficulty to manage the population changes over time or different overlapping populations. Usually, to overcome this problem, an overall population is constructed as a union of all units that ever existed, or as a union of different overlapping populations.

[1] Anton Grafström, Department of Forest Resource Management, Swedish University of Agricultural Sciences, SE-90183, Umea, Sweden. E-mail: anton.grafstrom@slu.se; Alina Matei, Institute of Statistics, University of Neuchâtel, Rue Bellevaux 51, 2000, Neuchâtel and Institute of Pedagogical Research and Documentation (IRDP) Neuchâtel, Switzerland. E-mail: alina.matei@unine.ch.

Various methods to provide sample coordination have been introduced in the literature. A summary of such methods is given for instance in Grafström and Matei (2015). An easy method to provide sample coordination is based on the use of so-called permanent random numbers introduced by Brewer, Early and Joyce (1972) for Poisson samples: one associates to each unit in the overall population an $U(0,1)$ random number. Such a number is called a permanent random number (PRN); these numbers are independent and are used in all sample selections. The probabilistic dependence of the samples' selection is thus created based on the use of permanent random numbers. Versions of the PRN method of Brewer et al. (1972) have been introduced in the literature (see Kröger, Särndal and Teikari, 1999; Kröger, Särndal and Teikari, 2003, for instance) and are widely used in different contexts. A recent example of a PRN method is the new system to coordinate business surveys by Statistics Canada. A two-phase stratified sampling design is used. The first-phase is a stratified sampling by Geography × Industry type × Business size and a Bernoulli sample is selected in each stratum by the use of PRNs. The main goal of the first-phase is to select a large sample covering all industries. For two consecutive first-phase waves a positive coordination is employed. In the second-phase, a sample is selected from the first-phase sample. For two consecutive second-phase waves, a negative coordination is applied to control the response burden of the business units (Haziza, 2013).

Our interest is to provide solutions to coordinate spatially balanced samples (for an overview on spatially balanced samples see Benedetti, Piersimoni and Postiglione, 2017). Usually, spatial sampling uses a space discretization, leading to the use of the classical sampling definition for finite populations. Thus, a population is defined as a finite set of units or locations having associated geographical coordinates. In most of the cases data are spatially autocorrelated and nearby locations tend to provide similar information. Consequently, it is desirable to sample units spread across the whole area of interest and to obtain a *spatially balanced sample*. The intuitive idea behind this is to cover through sampling the entire area of interest in order to obtain some representativeness. The selected sample should thus provide a full spatial coverage. Spatially balanced samples are efficient if a spatial trend is present in the variable of interest, denoted by y. Benedetti et al. (2017, page 447) note that "The motivation for the choice of selecting spatial well-spread samples is surely realistic if it is considered to be acceptable that increasing the distance between two units k and ℓ increases the difference, observed at units k and ℓ, namely, $|y_k - y_\ell|$. In this situation, it is evident that the variance of the Horvitz-Thompson estimator will necessarily decrease if we set high joint inclusion probabilities to pairs that have very different y values." Two spatial schemes useful for these goals are the local pivotal method (Grafström, Lundström and Schelin, 2012) and the spatially correlated Poisson sampling (Grafström, 2012). It was empirically found that both sampling schemes provide a good degree of spatial spreading, measured using Voronoi polytopes (see for instance Grafström et al., 2012, for some results).

We focus on coordination of spatially balanced samples using PRN methods, where sample selection is ensured using the local pivotal method (LPM) and the spatially correlated Poisson sampling (SCPS). Spatial sampling is used in many applications in environmental studies, forestry, agricultural surveys, but also in official statistics. We motivate the introduction of the coordinated spatially balanced samples by giving the following examples:

- In ecological monitoring, it is important to preserve over time the same sampled spatial locations in order to measure the changes in species abundance. However, over time, these locations may disappear. Positive coordination can be applied in this case to ensure a significant overlap of the selected locations.

- In national forest inventories, both current state and change of several parameters, such as growing stock volume for different tree species, are of interest. The methodology we present can be used to make sure the sample is continuously updated, e.g., yearly, to be well-spread geographically or in auxiliary variables available from remote sensing (to improve estimates of current state); a high positive coordination would guarantee good estimates of change as well.
- Different official national business registers contain spatial coordinates of business units (e.g., US Census Bureau's Longitudinal Business Database, the Swiss GeoStat, the Italian Statistical Archive of Active Enterprisers). The business units can be selected based on their geographical coordinates (Dickson, Benedetti, Giuliani and Espa, 2014). A negative coordination would be useful in this case to control the response burden of the units sampled by different surveys.

Note that methods to coordinate spatial samples have not yet been introduced in the literature. The novelty of the paper consists in introducing methods to coordinate spatially balanced samples. All the benefits of the sample coordination described above are provided for spatially balanced samples. In both types of coordination, the proposed methods preserve the spatial balancing property of the selected samples. Note that our goal is to control the overlap size between balanced samples, and not to improve sample coordination in general.

The paper is organized as follows. Section 2 introduces the notation. Sections 3.1 and 3.2 remind the local pivotal (LP) method and spatially correlated Poisson (SCP) sampling, respectively, while Section 3.3 a measure of spatial balance based on the Voronoi polytopes. We introduce methods to coordinate LP samples and SCP samples in Section 4. The same section introduces a new family of balanced sampling designs derived from SCP sampling, that provides good results for sample coordination. The coordination performances of the methods are presented in Section 5.1. Section 5.2 compares the new family of balanced sampling designs with Poisson sampling, while Section 5.3 provides simulation results for two typical estimators in repeated surveys. Section 6 shows an application of the proposed methods on real data. Discussion of the proposed methods and conclusions are provided in Section 7.

2 Notation

Let U_1 and U_2 be a population (subject to change over time) at time 1 and time 2, respectively, or consider that U_1 and U_2 are two overlapping populations. Consider samples s_1 and s_2 drawn from U_1 and U_2, using the sampling designs p_1 and p_2, respectively. No restriction about the sampling designs p_1 and p_2 is necessary to introduce the definitions in this section: they can be fixed or random size sampling designs, with or without replacement.

Let $U = U_1 \cup U_2$. We call U the "overall population". The set of labels of the units in U is $\{1, 2, \ldots, i, \ldots, N\}$. We define on U the joint sampling design p used to select a couple (s_1, s_2). The samples s_1 and s_2 are coordinated if $p(s_1, s_2) \neq p_1(s_1) p_2(s_2)$, that is the samples are not drawn independently (see Cotton and Hesse, 1992; Mach, Reiss and Şchiopu-Kratina, 2006). Let $\pi_{i1} = P(i \in s_1)$

and $\pi_{i2} = P(i \in s_2)$ be the first-order inclusion probabilities of unit $i \in U$ in the first and second sample, respectively. It follows that $\pi_{i1} = 0$ if $i \notin U_1$ and $\pi_{i2} = 0$ if $i \notin U_2$. Thus, it is not necessary to identify explicitly the subpopulation memberships.

Let $\pi_{i,12} = P(i \in s_1, i \in s_2)$ be the joint inclusion probability of unit $i \in U$ in both samples s_1 and s_2. If the samples s_1 and s_2 are selected independently, $\pi_{i,12} = \pi_{i1}\pi_{i2}$, for all $i \in U$.

Let c be the overlap between s_1 and s_2, which represents the number of common units of the two samples; it is in most of the cases a random variable. The coordination degree of s_1 and s_2 is measured by the expected overlap

$$E(c) = \sum_{i \in U} \pi_{i,12},$$

where $\pi_{i,12} = P(i \in s_1, i \in s_2)$. By using the Fréchet bounds of the joint probability $\pi_{i,12}$ it follows that

$$\sum_{i \in U} \max(0, \pi_{i1} + \pi_{i2} - 1) \leq E(c) = \sum_{i \in U} \pi_{i,12} \leq \sum_{i \in U} \min(\pi_{i1}, \pi_{i2}). \tag{2.1}$$

In negative coordination one wants to achieve the left bound in expression (2.1), that is $\sum_{i \in U} \max(0, \pi_{i1} + \pi_{i2} - 1) = E(c)$, while in positive coordination the right bound, that is $E(c) = \sum_{i \in U} \min(\pi_{i1}, \pi_{i2})$. Thus, to optimize the sample coordination process, the goal is to achieve these bounds, prior to coordination type, positive or negative. Using the terminology of Matei and Tillé (2005) the left side-part in (2.1) is called the Absolute Lower Bound (ALB) and the right side-part in (2.1) the Absolute Upper Bound (AUB).

The focus here is on sample coordination using PRNs. The PRN method was originally introduced by Brewer et al. (1972) to coordinate Poisson samples. Poisson sampling with PRNs reaches the Fréchet bounds given in equation (2.1). Yet, it results in a random sample size and does not provide spatially balanced samples. In order to achieve spatial balance, the local pivotal method (Grafström et al., 2012) and the spatially correlated Poisson sampling (Grafström, 2012) are used. Both sampling designs provide a good degree of spatial balance (see Grafström et al., 2012, for some empirical results). Moreover, since both are fixed size πps sampling designs (probability proportional to size sampling, see Särndal, Swensson and Wretman, 1992, page 90), the precision of the estimators is in general improved compared to Poisson sampling.

In what follows, we consider the sampling designs p_1 and p_2 to be without replacement, and the expected sample sizes of s_1 and s_2 are denoted by n_1 and n_2, respectively.

3 Spatial balanced sampling

The two spatial sampling designs we intend to introduce coordination for are briefly recalled below for a generic sample s of fixed size n.

3.1 Local pivotal method

The local pivotal method (Grafström et al., 2012) is a spatial application of the pivotal method (Deville and Tillé, 1998). Let $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_N)$ be a given vector of inclusion probabilities, with sum n,

$\pi_i = P(i \in s), i \in U.$ The vector $\boldsymbol{\pi}$ is successively updated to become a vector with $N - n$ zeros and n ones, where the ones indicate the selected units. A unit that still has a (possibly updated) probability strictly between 0 and 1 is called *undecided*. In one step of the LPM, a pair of units $i, j \in U$ is chosen to compete. More precisely, we choose unit i randomly among the undecided units, and unit i's competitor j is the nearest neighbor of i among the undecided units. Thus we apply the pivotal method locally in space. The winner receives as much probability mass as possible from the loser, so the winner ends up with $\pi_w = \min(1, \pi_i + \pi_j)$ and the loser keeps what is possibly remaining $\pi_\ell = \pi_i + \pi_j - \pi_w.$ The rules of the competition are

$$(\pi_i, \pi_j) := \begin{cases} (\pi_w, \pi_\ell) & \text{with probability } (\pi_w - \pi_j)/(\pi_w - \pi_\ell) \\ (\pi_\ell, \pi_w) & \text{with probability } (\pi_w - \pi_i)/(\pi_w - \pi_\ell) \end{cases}. \tag{3.1}$$

The final outcome is decided for at least one unit each update, so the procedure has at most N steps. Because neighboring units compete against each other for inclusion, they are unlikely to be simultaneously included in a sample.

3.2 Spatially correlated Poisson sampling

The spatially correlated Poisson sampling method (Grafström, 2012) is a spatial application of the correlated Poisson sampling method (Bondesson and Thorburn, 2008). Let $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_N)$ be a given vector of inclusion probabilities, with sum $n,$ $\pi_i = P(i \in s), i \in U.$ The vector $\boldsymbol{\pi}$ is sequentially updated to become a vector with $N - n$ zeros and n ones, where the ones indicate the selected units. First unit 1 is included with probability $\pi_1^{(0)} = \pi_1.$ If unit 1 was included, we set $I_1 = 1$ and otherwise $I_1 = 0.$ Generally at step $j,$ when the values for $I_1, \ldots, I_{j-1}$ have been recorded, unit j is included with probability $\pi_j^{(j-1)}.$ Then the inclusion probabilities are updated for the units $i = j + 1, \ldots, N,$ according to

$$\pi_i^{(j)} = \pi_i^{(j-1)} - \left(I_j - \pi_j^{(j-1)}\right) w_j^{(i)}, \tag{3.2}$$

where $w_j^{(i)}$ are weights given by unit j to the units $i = j + 1, j + 2, \ldots, N$ and $\pi_i^{(0)} = \pi_i.$ The weight $w_j^{(i)},$ $j < i,$ determine how the inclusion probability for unit i should be affected by the sampling outcome of unit $j.$ More precisely, the weight $w_j^{(i)},$ $j < i,$ may depend on the previous sampling outcome $I_1, I_2, \ldots, I_{j-1}$ but not on the future outcomes $I_j, I_{j+1}, \ldots, I_N.$ The weights should also satisfy the following restrictions

$$-\min\left(\frac{1 - \pi_i^{(j-1)}}{1 - \pi_j^{(j-1)}}, \frac{\pi_i^{(j-1)}}{\pi_j^{(j-1)}}\right) \le w_j^{(i)} \le \min\left(\frac{\pi_i^{(j-1)}}{1 - \pi_j^{(j-1)}}, \frac{1 - \pi_i^{(j-1)}}{\pi_j^{(j-1)}}\right)$$

in order for $0 \le \pi_i^{(j-1)} \le 1,$ $i = j, j + 1, \ldots, N,$ to hold. The unconditional inclusion probabilities are not affected by the weights since the updating rule (3.2) gives

$$E\left(\pi_i^{(i-1)}\right) = E\left(E\left(\pi_i^{(i-1)} \middle| \pi_i^{(i-2)}\right)\right) = E\left(\pi_i^{(i-2)}\right) = \ldots = \pi_i.$$

Thus the method always gives the prescribed inclusion probabilities $\pi_i, i = 1, 2, \ldots, N.$

Bondesson and Thorburn (2008) showed that a fixed size sampling is obtained only if $\sum_{i=1}^N \pi_i = n$ and the weights are chosen such that $\sum_{i=j+1}^N w_j^{(i)} = 1,$ $j \in U.$

To achieve spatial balance, the weights should be decided on the basis of the distance between units. The most common approach to choose weights in SCPS is that unit j first gives as much weight as possible to the closest unit (in distance) among the units $i = j+1, j+2, \ldots, N,$ then as much weight as possible to the second closest unit etc. with the restriction that the weights are non-negative and sum up to 1. This strategy is called the *maximal weight strategy*. If distances are equal, then the weight is distributed equally on those units that have equal distance if possible. The first priority is that weight is not put on a unit if it is possible to put the weight on a closer unit. The maximal weight strategy always produces samples of fixed size if the inclusion probabilities sum up to an integer. In what follows, when we refer to SCPS, the "maximal weight strategy" is used.

3.3 Voronoi polytopes

Voronoi polytopes are used to measure the level of spatial balance (or spread) with respect to the inclusion probabilities (Stevens and Olsen, 2004). A polytope P_i is constructed for each unit $i \in s,$ and P_i includes all population units closer to unit i than to any other sample unit $j \in s, j \neq i.$ Optimally, each polytope should have a probability mass that is equal to 1. A measure of spatial balance of a realised sample s is (see Stevens and Olsen, 2004)

$$B = \frac{1}{n}\sum_{i \in s}(v_i - 1)^2, \tag{3.3}$$

where v_i is the sum of the inclusion probabilities of the units in P_i. The expected value of B under repeated sampling is a measure of how well a design succeeds in selecting spatially balanced samples. The smaller the value the better the spread of the selected samples.

4 Coordination methods

We present below PRN methods based on the local pivotal (LP) method and the spatially correlated Poisson (SCP) sampling.

4.1 Coordination of LP samples

The positive coordination of LP samples with PRNs is implemented as follows:

1. independent permanent random numbers $v_{ij} \sim U(0,1)$ are associated to each pair $(i, j) \subseteq U \times U;$
2. s_1 is drawn using LPM as follows: if a pair of units (i, j) is chosen to compete, the number v_{ij} is used in the corresponding competition rule (3.1) and the pair (i, j) is saved into a list of pairs;
3. s_2 is drawn using LPM as follows: the pairs (i, j) are considered sequentially from the list of pairs constructed above for s_1, and the same numbers v_{ij} are used in the corresponding competition rule (3.1). If the sample size n_2 is achieved using pairs from this list, the algorithm stops; if not, the selection process continues with new pairs (i, j) (not included in this list) and selected as described in Section 3.1.

For negative coordination, the first two steps are the same, but the last step becomes:

3'. s_2 is drawn using LPM as follows: the pairs (i, j) are considered sequentially from the list of pairs constructed above for s_1, and the numbers $1 - v_{ij}$ are used in the corresponding competition rule (3.1). If the sample size n_2 is achieved using pairs from this list, the algorithm stops; if not, the selection process continues with new pairs (i, j) (not included in this list) and selected as described in Section 3.1.

4.2 Coordination of SCP samples

The coordination of SCP samples with PRNs is implemented as follows. Let u_i be the PRN associated to unit $i \in U$, with $u_1, u_2, \ldots, u_N$ iid $U(0, 1)$. Let $\pi_{it}^{(i-1)}$ be the (updated) selection probability for unit i in the selection of sample s_t, $t = 1, 2$. For positive coordination, the PRNs are introduced in the selection step similarly to Poisson sampling with PRNs: if $u_i < \pi_{it}^{(i-1)}$, unit i is selected in the sample s_t, $t = 1, 2$. For negative coordination, if $u_i < \pi_{i1}^{(i-1)}$, unit i is selected in s_1; if $1 - u_i < \pi_{i2}^{(i-1)}$, unit i is selected in s_2. This coordination method is general for spatially correlated Poisson sampling and can be used no matter what weights are applied within the method.

We utilize the maximal weight strategy advocated in Section 3.2 as the main alternative, but we also introduce two new alternative strategies to compute the weights $w_j^{(i)}$. The new strategies are intended to provide a good compromise between the degrees of spatial balance and coordination. By reducing the amount of spatial correlation in SCPS we can achieve any level of mixing between SCPS and Poisson sampling. Both of the new strategies are similar to the SCPS with maximal weights, but the weights $w_j^{(i)}$ given by the unit j to units $i = j + 1, \ldots, N$ do not sum up to 1 any more. Consequently, the result of Bondesson and Thorburn (2008) advocated in Section 3.2 does not apply and the new sampling designs do not any more provide fixed sample sizes. We denote the resulting family of designs Transformed Spatially Correlated Poisson Sampling (TSCPS).

The first mixing strategy is to modify SCPS by multiplying the maximal weight by a given scalar α, $0 \le \alpha \le 1$. Thus we no longer use maximal weight, but the proportion α of the maximal weight is the limit for the applied weight. This method is denoted TSCPS 1. With this method the positive weights will reach longer (more neighbors) than in SCPS. Each unit would distribute a total weight of maximum 1, starting with the nearest unit and then the second nearest etc. Say the maximal weights for the three nearest neighbors of a unit are 0.7, 0.5, 0.2. Then, in standard SCPS (with maximal weights) the unit would distribute the weights 0.7, 0.3, 0, and the new modified version would, with $\alpha =$ 0.5, distribute the weights 0.35, 0.25, 0.1. The reach is longer but it is not guaranteed we can use all α. As a result, the total weight is not necessary 1, and the sample size becomes random.

The second mixing strategy is achieved by limiting the weights that a unit distributes to sum to a fixed scalar α, $0 \le \alpha \le 1$. This method is denoted TSCPS 2. In SCPS with maximal weight strategy, each unit is given a total weight 1 (the sum of the weights) to distribute on remaining units in the list. Instead, each unit is given a total weight α to distribute. Otherwise, this works as the maximal weight strategy, so that unit i first gives as much weight as possible to the nearest, then the second nearest etc. With this strategy the weights will reach a shorter distance (fewer neighbors). Say the maximal weights for the three nearest neighbors of a unit are 0.7, 0.5, 0.2. Then standard SCPS (with maximal weights) would distribute the weights 0.7, 0.3, 0, and the new modified version would, with $\alpha =$ 0.5, distribute 0.5, 0, 0. The reach is shorter and it is guaranteed we can use all α. However, if the total weight α is less than 1, there will be a random sample size.

Note that for both TSCPS 1 and 2 we have the following result. With $\alpha = 0,$ we get Poisson sampling and with $\alpha = 1$ we get SCPS with maximal weights. We can scale with α between 0 and 1 to mix the two to any degree. Maximum coordination, worst spatial balance and highest variance of sample size for $\alpha = 0,$ and best spatial balance and guaranteed fixed sample size for $\alpha = 1$ while level of coordination will be to some extent worse. Both TSCPS 1 and 2 offer the possibility to make a trade-off between the Poisson and SCPS designs. Degree of spatial balance and coordination, as well as variance of achieved sample size depend on the parameter α. Sample size is likely to be more stable (given the same α) for TSCPS 1 than for TSCPS 2, as more weight is likely to be distributed with TSCPS 1. Since both TSCPS 1 and 2 use a given scalar $\alpha,$ $0 \le \alpha \le 1,$ they provide a family of sampling designs. Each element in this family corresponds to a given α. Contrary to SCPS, for any value of $\alpha < 1$ both TSCPS 1 and 2 involve random sample sizes. The consequences of having random sample sizes on coordination is empirically studied in Section 5.1, on spatial balance degree in Section 5.2 and on variance estimation in Section 5.3.

5 Empirical results

5.1 Overlap performance

Monte Carlo simulation was used to study the overlap performance of the proposed methods. A number of $m = 10^4$ runs were considered for each of the four settings described below. In each run, samples were drawn using the proposed methods. The same permanent random numbers were employed for all methods. The Euclidean distance between units was used for all spatial sampling designs. In each run, for LPM with PRNs, a matrix of dimension $N \times N$ of PRNs was randomly generated; the diagonal elements of this matrix were used as PRNs for Poisson, SCPS and the transformed SCPS with PRNs. All sampling schemes were applied for positive and negative coordination, respectively, using in each run the same PRNs and the same matrix of distances. Samples s_1 and s_2 of following types were drawn in each run:

- two Poisson samples selected respectively independently, positively coordinated with PRNs, and negatively coordinated with PRNs;
- two LP samples selected respectively independently, positively coordinated with PRNs, and negatively coordinated with PRNs;
- two SCP samples selected respectively independently, positively coordinated with PRNs, and negatively coordinated with PRNs;
- two transformed SCP samples selected respectively independently, positively coordinated with PRNs, and negatively coordinated with PRNs; the two strategies shown in Section 4.2 were employed using respectively $\alpha =$ 0.25, 0.50 and 0.75.

Three measures were used to quantify the performance of the proposed methods, for positive and negative coordination, respectively:

- the Monte Carlo expected overlap

$$E_{\text{sim}}(c) = \frac{1}{m}\sum_{\ell=1}^{m} c_{\ell}^{1,2},$$

  $c_{\ell}^{1,2} = |s_{1\ell} \cap s_{2\ell}|,$ and $s_{1\ell}, s_{2\ell},$ are the samples drawn in the ℓ^{th} run, where $|s_{1\ell} \cap s_{2\ell}|$ represents the number of common units of $s_{1\ell}$ and $s_{2\ell}$;
- the Monte Carlo variance of the overlap

$$V_{\text{sim}}(c) = \frac{1}{m-1}\sum_{\ell=1}^{m}\left(c_{\ell}^{1,2} - E_{\text{sim}}(c)\right)^2;$$

- the Monte Carlo coefficient of variation of the overlap

$$\text{CV}_{\text{sim}}(c) = \frac{\sqrt{V_{\text{sim}}(c)}}{E_{\text{sim}}(c)}.$$

The correlation between $\boldsymbol{\pi}_1 = (\pi_{1i})_{i=1,\ldots,N}$ and $\boldsymbol{\pi}_2 = (\pi_{2i})_{i=1,\ldots,N}$ is an important factor of the sample coordination degree. This correlation varies and takes extreme values in the following four settings used to study the performance of the proposed methods:

- the static MU284 population: from the MU284 data set (see Appendix B in Särndal et al., 1992), the region 2 was selected. The population size is $N = 48,$ and the expected sample sizes are $n_1 = 10, n_2 = 6,$ respectively. The first-order inclusion probabilities π_{i1} are computed using the variable P75 (population in 1975 in thousands), and π_{i2} using the variable P85 (population in 1985 in thousands). The elements of the distance matrix were artificially generated using independent draws from the $N(0,1)$ distribution and taking their absolute values. The correlation coefficient between $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ is 0.99.
- the Baltimore data set is about house sales prices and hedonics (see Dubin, 1992). The data set is available on-line at the GeoDa Center for Geospatial Analysis and Computation (2017). Information on $N = 211$ houses are provided by 17 variables. The geographical coordinates of the houses are available. We use $n_1 = n_2 = 25.$ The first-order inclusion probabilities π_{i1} are computed using the variable AGE (the house age) and π_{i2} using AGE+5. The elements of the distance matrix are the Euclidean distances between the geographical coordinates on the Maryland grid of the houses included in this data set. The correlation coefficient between $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ is 1.
- the MU284 dynamic population: from the MU284 data set, the regions 2 and 3 were used. A dynamic population was created using on the first occasion 50% of the units randomly selected from the region 2 using simple random sampling without replacement (these units are the "persistents" and the rest of the 50% of the units are "deaths"), and on the second occasion 50% of the units randomly selected from the region 3 using simple random sampling without replacement (these units are the "births"). The elements of the distance matrix were artificially

generated using independent draws from the $N(0,1)$ distribution and taking their absolute values. For a run, the correlation coefficient between $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ was 0.08.

- one artificial data set, with $N = 100$, $n_1 = 10$, $n_2 = 25$, $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ uncorrelated and randomly generated using independent draws from the $U(0,1)$ distribution and scaled to obtain the sum 10 and 25, respectively. The elements of the distance matrix were artificially generated using independent draws from the $N(0,1)$ distribution and taking their absolute values.

A number of 10^4 simulation runs was used to compute the Monte Carlo overlap measures using the nine methods in each setting. Tables 5.1, 5.2, 5.3, and 5.4 provide the results of the Monte Carlo studies based on the previous four settings. For TSCPS 1 and 2, the value of α is also specified in these tables.

Table 5.1
The static MU284 population, $N = 48$, expected sample sizes $n_1 = 10, n_2 = 6$, π_{i1} are computed using the variable P75 (population in 1975 in thousands), and π_{i2} using the variable P85 (population in 1985 in thousands). The distance matrix was artificially generated. The values of AUB and ALB are 6 and 1.96, respectively

Method		independent			positive			negative		
		$E_{\text{sim}}(c)$	$V_{\text{sim}}(c)$	$\text{CV}_{\text{sim}}(c)$	$E_{\text{sim}}(c)$	$V_{\text{sim}}(c)$	$\text{CV}_{\text{sim}}(c)$	$E_{\text{sim}}(c)$	$V_{\text{sim}}(c)$	$\text{CV}_{\text{sim}}(c)$
Poisson		3.04	1.89	0.45	6.03	4.06	0.33	1.96	1.13	0.54
LPM		3.03	1.22	0.36	5.10	0.71	0.17	2.64	1.20	0.41
SCPS		3.06	1.21	0.36	4.91	0.85	0.19	2.33	1.06	0.44
TSCPS 1	$\alpha =$ 0.25	3.06	1.28	0.37	5.84	0.93	0.17	2.09	1.13	0.51
	$\alpha =$ 0.50	3.04	1.27	0.37	5.54	0.79	0.16	2.21	1.10	0.47
	$\alpha =$ 0.75	3.06	1.25	0.37	5.20	0.80	0.17	2.27	1.06	0.45
TSCPS 2	$\alpha =$ 0.25	3.07	1.67	0.42	5.75	2.40	0.27	1.97	1.13	0.54
	$\alpha =$ 0.50	3.06	1.45	0.39	5.40	1.57	0.23	2.05	1.10	0.51
	$\alpha =$ 0.75	3.04	1.27	0.37	5.13	1.10	0.20	2.18	1.04	0.47

Table 5.2
Baltimore data, $N = 211$, expected sample sizes $n_1 = 25, n_2 = 25$, π_{i1} are computed using the variable AGE and π_{i2} using AGE+5. The distance matrix uses real data. The values of AUB and ALB are 24.20 and 0.10, respectively

Method		independent			positive			negative		
		$E_{\text{sim}}(c)$	$V_{\text{sim}}(c)$	$\text{CV}_{\text{sim}}(c)$	$E_{\text{sim}}(c)$	$V_{\text{sim}}(c)$	$\text{CV}_{\text{sim}}(c)$	$E_{\text{sim}}(c)$	$V_{\text{sim}}(c)$	$\text{CV}_{\text{sim}}(c)$
Poisson		4.08	3.93	0.49	24.20	20.63	0.19	0.10	0.09	3.00
LPM		4.09	3.15	0.43	21.50	2.86	0.08	1.76	1.51	0.70
SCPS		4.01	3.22	0.45	22.20	3.14	0.08	0.76	0.70	1.10
TSCPS 1	$\alpha =$ 0.25	4.05	3.02	0.43	23.10	2.60	0.07	0.26	0.26	1.96
	$\alpha =$ 0.50	4.06	3.06	0.43	22.50	2.93	0.08	0.45	0.43	1.46
	$\alpha =$ 0.75	4.05	3.22	0.44	22.30	3.10	0.08	0.57	0.55	1.30
TSCPS 2	$\alpha =$ 0.25	4.07	3.56	0.46	23.70	11.75	0.14	0.10	0.09	3.00
	$\alpha =$ 0.50	4.07	3.37	0.45	23.20	6.35	0.11	0.29	0.27	1.79
	$\alpha =$ 0.75	4.04	3.31	0.45	22.70	3.84	0.09	0.58	0.52	1.24

Table 5.3
The dynamic MU284 population – region 2 from the MU284 population, where 50% of the units are new in the second occasion (“births”), and 50% of the units change the stratum (“deaths”), N = 72, expected sample sizes n_1 = 10, n_2 = 6. The distance matrix was artificially generated. The values of AUB and ALB are 3.56 and 1.33, respectively

Method		independent			positive			negative		
		$E_{\text{sim}}(c)$	$V_{\text{sim}}(c)$	$\text{CV}_{\text{sim}}(c)$	$E_{\text{sim}}(c)$	$V_{\text{sim}}(c)$	$\text{CV}_{\text{sim}}(c)$	$E_{\text{sim}}(c)$	$V_{\text{sim}}(c)$	$\text{CV}_{\text{sim}}(c)$
Poisson		2.02	1.20	0.54	3.56	2.35	0.43	1.32	0.71	0.64
LPM		2.03	0.95	0.48	2.37	1.00	0.42	1.87	0.89	0.50
SCPS		2.02	1.02	0.50	3.01	1.19	0.36	1.54	0.79	0.58
TSCPS 1	α = 0.25	2.02	0.94	0.48	3.42	1.31	0.33	1.39	0.70	0.60
	α = 0.50	2.03	1.02	0.50	3.27	1.33	0.35	1.42	0.79	0.63
	α = 0.75	2.02	1.02	0.50	3.16	1.26	0.36	1.47	0.80	0.61
TSCPS 2	α = 0.25	2.02	1.04	0.50	3.36	1.67	0.38	1.33	0.64	0.60
	α = 0.50	2.02	0.96	0.49	3.20	1.37	0.37	1.41	0.66	0.58
	α = 0.75	2.02	0.94	0.48	3.10	1.24	0.36	1.50	0.71	0.56

Table 5.4
Artificial data, $N = 100$, expected sample sizes $n_1 = 10, n_2 = 25$, π_{i1} and π_{i2} randomly generated, uncorrelated. The distance matrix was artificially generated. The values of AUB and ALB are 9.11 and 0, respectively

Method		independent			positive			negative		
		$E_{\text{sim}}(c)$	$V_{\text{sim}}(c)$	$\text{CV}_{\text{sim}}(c)$	$E_{\text{sim}}(c)$	$V_{\text{sim}}(c)$	$\text{CV}_{\text{sim}}(c)$	$E_{\text{sim}}(c)$	$V_{\text{sim}}(c)$	$\text{CV}_{\text{sim}}(c)$
Poisson		2.44	2.34	0.63	9.11	8.08	0.31	~ 0	~ 0	
LPM		2.45	1.82	0.55	5.42	2.35	0.28	1.03	0.91	0.93
SCPS		2.42	1.82	0.56	6.94	2.07	0.21	0.45	0.42	1.44
TSCPS 1	α = 0.25	2.44	1.76	0.54	8.53	2.05	0.17	0.06	0.07	4.41
	α = 0.50	2.46	1.79	0.54	7.95	1.90	0.17	0.21	0.22	2.23
	α = 0.75	2.43	1.80	0.55	7.40	1.97	0.19	0.31	0.31	1.80
TSCPS 2	α = 0.25	2.43	2.09	0.59	8.53	4.86	0.26	~ 0	~ 0	
	α = 0.50	2.45	1.91	0.56	7.90	3.32	0.23	0.11	0.10	2.87
	α = 0.75	2.44	1.83	0.55	7.34	2.51	0.22	0.28	0.26	1.82

Following the results given in Tables 5.1, 5.2, 5.3, and 5.4, SCPS shows in general better performance than LPM in terms of $E_{\text{sim}}(c), V_{\text{sim}}(c)$ and $\text{CV}_{\text{sim}}(c)$ for both types of coordination; an exception is the case of the static MU284 population and positive coordination. In this setting, the pairs used for the selection of s_1 are also used for the selection of s_2, since deaths or births are not assumed. Without such changes in population, LPM may perform better than SCPS in terms of $E_{\text{sim}}(c)$, but also in terms of $V_{\text{sim}}(c)$ and $\text{CV}_{\text{sim}}(c)$.

As expected, Poisson sampling achieves the AUB and ALB (minor differences are due to the sampling error) in all settings, but the overlap variance is very high in positive coordination. This is mainly due to the random sizes of s_1 and s_2. The large values of $V_{\text{sim}}(c)$ impact the values of $\text{CV}_{\text{sim}}(c)$. In all the examples shown, the latter is in general larger than the values of $\text{CV}_{\text{sim}}(c)$ provided by the other sampling schemes.

Results in Tables 5.1, 5.2, 5.3, and 5.4 confirm that the value of α in the transformed SCPS determines the coordination degree; a smaller value of α provides a better coordination degree, since one gets closer to Poisson sampling (we remind that $\alpha = 0$ in the TSCPS designs leads to Poisson sampling).

For a given α, the new strategies presented in Section 4.2 yield similar values of $E_{\text{sim}}(c)$ in positive coordination, but TSCPS 2 gives larger values of $V_{\text{sim}}(c)$ and $\text{CV}_{\text{sim}}(c)$. For all α used, both TSCPS 1 and TSCPS 2 provides similar values of $\text{CV}_{\text{sim}}(c)$ in positive and negative coordination in our examples, excepting TSCPS 2 with $\alpha = 0.25$. The latter performs very close to Poisson sampling in negative coordination as the results in Tables 5.1, 5.2, 5.3, and 5.4 show.

An interesting result for Poisson sampling arises from Tables 5.1, 5.2, 5.3, and 5.4 in terms of $\text{CV}_{\text{sim}}(c)$. While the values of $V_{\text{sim}}(c)$ are large for positive coordination compared to LPM and SCPS, it is not the case for negative coordination. However, in the latter case, if $E_{\text{sim}}(c) \sim V_{\text{sim}}(c)$ and both are small as in Table 5.2, the corresponding value of $\text{CV}_{\text{sim}}(c)$ becomes very large. As we mentioned, that can also be the case for the TSCPS designs with small values of α. The improvement of introducing this new family of designs compared to Poisson sampling is measured for these situations in terms of spatial balance degree as shown in the next section.

5.2 Spatial balance and variance of sample size

The transformed SCPS is compared to the other sampling designs in terms of degree of spatial balance using Monte-Carlo simulation. The degree of spatial balance is measured using the B measure shown in expression (3.3). For the transformed SCPS the two strategies presented in Section 4.2 are used, and the four previous settings are employed. The B measure was computed on the same samples s_1 used to obtain the outcomes given in Tables 5.1, 5.2, 5.3, and 5.4, respectively. The following overall measure was used for each type of sample

$$E_{\text{sim}}(B) = \frac{1}{m} \sum_{\ell=1}^{m} B_{\ell},$$

where B_{ℓ} represents the B measure computed on a realised sample in the ℓ^{th} run. For comparison, the average of the B measures computed over the Monte-Carlo runs for Poisson sampling and LPM were also reported.

TSCPS is also compared with Poisson sampling in terms of variance of sample size computed over the Monte-Carlo runs using:

$$V_{\text{sim}}(\text{size}) = \frac{1}{m-1} \sum_{\ell=1}^{m} (|s_{\ell}| - |\bar{s}|)^2,$$

where $|s_{\ell}|$ represents the sample size of a realised sample s_{ℓ} in the ℓ^{th} run and $|\bar{s}| = \frac{1}{m} \sum_{\ell=1}^{m} |s_{\ell}|$.

Tables 5.5, 5.6, 5.7 and 5.8 provide the results. Following these results, we note that the choice of α determines the performance of the transformed SCPS in terms of averaged B measure: a larger value of α results in a better spatial balance degree. However, in all settings, the resulting spatial balance degree is worse than for LPM and SCPS, but better than for Poisson sampling as expected, since the latter is not a spatial balanced sampling.

For all four settings, the variance of sample size is much higher for Poisson sampling than for TSCPS 1 and TSCPS 2, for all values of α. While TSCPS 2 with $\alpha = 0.25$ performs very close to Poisson sampling in the examples shown in Section 5.1 for negative coordination, we note however that the corresponding values of $V_{\text{sim}}\,(\text{size})$ for the former method are much smaller than those provided by Poisson sampling.

As underlined in Section 4.2, TSCPS 1 shows smaller sample size variance than TSCPS 2 for the same α. The results in our settings confirm for both TSCPS 1 and TSCPS 2 that the variance of sample size decreases when α increases.

Table 5.5
The static MU284 population, $N = 48$, expected sample size 10, the inclusion prob. are computed using the variable P75 (population in 1975 in thousands). The distance matrix was artificially generated

Design		$E_{\text{sim}}\,(B)$	$V_{\text{sim}}\,(\text{size})$
Poisson		0.301	4.806
LPM		0.124	0
SCPS		0.131	0
TSCPS 1	$\alpha = 0.25$	0.209	0.727
	$\alpha = 0.50$	0.177	0.405
	$\alpha = 0.75$	0.146	0.187
TSCPS 2	$\alpha = 0.25$	0.215	2.692
	$\alpha = 0.50$	0.159	1.211
	$\alpha = 0.75$	0.134	0.399

Table 5.6
Baltimore data, $N = 211$, expected sample size 25, the inclusion prob. are computed using the variable AGE. The distance matrix uses real data

Design		$E_{\text{sim}}\,(B)$	$V_{\text{sim}}\,(\text{size})$
Poisson		0.416	21.107
LPM		0.137	0
SCPS		0.137	0
TSCPS 1	$\alpha = 0.25$	0.256	0.909
	$\alpha = 0.50$	0.198	0.449
	$\alpha = 0.75$	0.162	0.222
TSCPS 2	$\alpha = 0.25$	0.282	11.382
	$\alpha = 0.50$	0.195	4.811
	$\alpha = 0.75$	0.148	1.227

Table 5.7
The dynamic MU284 population, $N = 48,$ expected sample size 10, the inclusion prob. are computed using the variable P75 (population in 1975 in thousands). The distance matrix was artificially generated

Design		$E_{\text{sim}}(B)$	$V_{\text{sim}}(\text{size})$
Poisson		0.422	5.683
LPM		0.202	0
SCPS		0.210	0
TSCPS 1	$\alpha = 0.25$	0.306	0.798
	$\alpha = 0.50$	0.255	0.427
	$\alpha = 0.75$	0.224	0.231
TSCPS 2	$\alpha = 0.25$	0.315	3.128
	$\alpha = 0.50$	0.252	1.370
	$\alpha = 0.75$	0.213	0.446

Table 5.8
Artificial data, $N = 100,$ expected sample size 10, the inclusion prob. are randomly generated. The distance matrix was artificially generated

Design		$E_{\text{sim}}(B)$	$V_{\text{sim}}(\text{size})$
Poisson		0.485	8.892
LPM		0.134	0
SCPS		0.133	0
TSCPS 1	$\alpha = 0.25$	0.286	0.938
	$\alpha = 0.50$	0.213	0.446
	$\alpha = 0.75$	0.167	0.230
TSCPS 2	$\alpha = 0.25$	0.313	4.854
	$\alpha = 0.50$	0.204	2.121
	$\alpha = 0.75$	0.149	0.632

5.3 Variance estimation

In repeated surveys, estimates of net variation, period averages and gross change are of interest. Our proposed methods are suitable to estimate such parameters. Their variance estimation is, however, intractable for our methods and is not addressed here. We study only empirically the impact that each coordinated spatial balancing method has on the quality of the estimates of two of the above parameters. Note that there exist approximative variance estimators for state that can be used for LPM and SCPS (Grafström and Schelin, 2014), but further research is needed to derive an approximative estimator for the covariance between successive state estimators under coordination.

Consider a repeated survey over two time occasions. Let y be the variable of interest, measured in the first and second time occasion, respectively. We denote by y_{it} the value of this variable taken by the unit $i \in U$ on the time occasion t, with $t \in \{1, 2\}$. Let x_{it} be the value of an auxiliary variable taken by the unit $i \in U$ at occasion t; the variable x is well correlated with y, and available for all units $i \in U$ in both time occasions. It is assumed that x_{it} is known for all $i \in U$ from a previous census or that a two-phase sampling is used: in the first phase the value of x_{it} is obtained, while the coordination process is addressed in the second phase of the sampling. The notation $E_M\,(.)$ and $\operatorname{var}_M\,(.)$ indicate the expectation and variance under a model. We borrow from Grafström and Tillé (2013) the following cross-sectional superpopulation model with spatial correlation

$$y_{i,\,t-1} \;=\; \beta_0 + x_{i,\,t-1}\beta_1 + \varepsilon_{i,\,t-1}, \tag{5.1}$$

where β_0 and β_1 are parameters, where $\varepsilon_{i,\,t-1}$ are random variables, with $E_M\left(\varepsilon_{i,\,t-1}\right) = 0,$ $\operatorname{var}_M\left(\varepsilon_{i,\,t-1}\right) = \sigma_i^2,$ $\operatorname{cov}_M\left(\varepsilon_i, \varepsilon_j\right) = \sigma_i\sigma_j\rho^{d(i,\,j)},$ where $d(i,\, j)$ represents the distance between the units i and $j,$ for $i,\, j \in U.$ The particular form of $\operatorname{cov}_M\left(\varepsilon_i, \varepsilon_j\right)$ in model (5.1) underlines a decreasing function of the distance between i and $j,$ reflecting that the proximity of units implies a larger spatial correlation. The following autoregressive model is considered

$$y_{it} \;=\; \delta_0 + \delta_1 y_{i,\,t-1} + \gamma_{it}, \tag{5.2}$$

with δ_0 and δ_1 being parameters, and with γ_{it} being independent random variables, with $E_M\,(\gamma_{it}) =$ $0,$ $\operatorname{var}_M\,(\gamma_{it}) = u^2.$ The following model is also assumed

$$x_{it} \;=\; \alpha_0 + \alpha_1 x_{i,\,t-1} + \tilde{\gamma}_{it}, \tag{5.3}$$

where α_0 and α_1 are parameters, where $\tilde{\gamma}_{it}$ are independent random variables, with $E_M\,(\tilde{\gamma}_{it}) = 0,$ $\operatorname{var}_M\,(\tilde{\gamma}_{it}) = \tilde{u}^2.$ We obtain thus a spatial-temporal dependence of the data through models (5.1), (5.2) and (5.3).

We consider that π_{it} are constructed using the expression

$$\pi_{it} \;=\; \frac{n_t x_{it}}{\sum_{j\in U} x_{jt}},\; t \in \{1,\, 2\},$$

that leads to a correlation between $\boldsymbol{\pi}_{t-1}$ and $\boldsymbol{\pi}_t$ due to model (5.3).

The following parameters of interest are considered: the one period change $D = \sum_{i\in U_1} y_{i_1} - \sum_{i\in U_2} y_{i2}$ and the average over two periods $A = \frac{1}{2}\left(\sum_{i\in U_1} y_{i_1} + \sum_{i\in U_2} y_{i2}\right).$ The two parameters are estimated by

$$\hat{D} \;=\; \sum_{i\in s_1}\frac{y_{i_1}}{\pi_{i1}} - \sum_{i\in s_2}\frac{y_{i2}}{\pi_{i2}},$$

and

$$\hat{A} \;=\; \frac{1}{2}\left(\sum_{i\in s_1}\frac{y_{i_1}}{\pi_{i1}} \;+\; \sum_{i\in s_2}\frac{y_{i2}}{\pi_{i2}}\right),$$

respectively. We have

$$\operatorname{var}\left(\hat{D}\right) \;=\; \operatorname{var}\left(\sum_{i\in s_1}\frac{y_{i_1}}{\pi_{i1}}\right) \;+\; \operatorname{var}\left(\sum_{i\in s_2}\frac{y_{i2}}{\pi_{i2}}\right) \;-\; 2\operatorname{cov}\left(\sum_{i\in s_1}\frac{y_{i_1}}{\pi_{i1}}, \sum_{i\in s_2}\frac{y_{i2}}{\pi_{i2}}\right), \tag{5.4}$$

$$\operatorname{var}\left(\hat{A}\right) \;=\; \frac{1}{4}\operatorname{var}\left(\sum_{i\in s_1}\frac{y_{i_1}}{\pi_{i1}}\right) \;+\; \frac{1}{4}\operatorname{var}\left(\sum_{i\in s_2}\frac{y_{i2}}{\pi_{i2}}\right) \;+\; \frac{1}{2}\operatorname{cov}\left(\sum_{i\in s_1}\frac{y_{i_1}}{\pi_{i1}}, \sum_{i\in s_2}\frac{y_{i2}}{\pi_{i2}}\right), \tag{5.5}$$

where $\operatorname{var}(.)$ and $\operatorname{cov}(.,.)$ represent the variance and the covariance operators, respectively.

Following expression (5.4), if s_1 and s_2 are positively coordinated, the variance of $\hat{D}$ is reduced in general through sample overlap, since a positive covariance between $\sum_{i \in s_1} y_{i_1} / \pi_{i1}$ and $\sum_{i \in s_2} y_{i2} / \pi_{i2}$ is achieved compared to independent samples' selection. Similarly, from expression (5.5), independent samples' selection reduces the variance of $\hat{A}$ compared to positively coordinated samples because this covariance is zero. Negative coordination of samples can lead to a negative covariance between $\sum_{i \in s_1} y_{i_1} / \pi_{i1}$ and $\sum_{i \in s_2} y_{i_2} / \pi_{i2}$, and the variance of $\hat{A}$ can diminish compared to independent samples' selection.

A population of size $N = 100$ was created using models (5.1), (5.2), and (5.3). No births or deaths were considered in the population. The distance matrix was artificially generated using absolute values of independent runs from the $N(0,1)$ distribution. We set $\beta_0 = 4$, $\beta_1 = 2$, $\rho = 0.9$, $\delta_0 = 0$, $\delta_1 = 1$, $\alpha_0 = 0$, $\alpha_1 = 1$, $\tilde{\gamma}_i \sim N(0,1), i = 1, \ldots, N,$ iid and $\gamma_i = \beta_1 \tilde{\gamma}_i, i = 1, \ldots, N.$ We also generated artificially x_{i1} as independent random draws from the $N(4,1)$ distribution. The correlation between y_1 and y_2 was approximately 0.72, while between y_t and $x_t, t = 1, 2$ was approximately 0.9. Based on this population, two different settings were created, by varying n_1 and n_2: in the first setting $n_1 = 10, n_2 = 25,$ while in the second one $n_1 = n_2 = 50.$ The correlation between $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ was approximately 0.7 in both settings.

Monte Carlo simulation was used to study empirically the impact that each proposed method has on $\operatorname{var}(\hat{D})$ and $\operatorname{var}(\hat{A})$. For each setting, $m = 10^4$ samples were drawn as described in the beginning of Section 5.1. Figures 5.1 and 5.2 show boxplots corresponding to the $\hat{D}$ values obtained through Monte Carlo simulation, for both settings. The white boxplots correspond to the $\hat{D}$ values obtained from independent samples s_1 and s_2, while the grey ones to positively coordinated samples s_1 and s_2. The sampling design is specified below each boxplot (for example, TSCPS1_indep_0.25 indicates TSCPS 1 with independent samples' selection and $\alpha =$ 0.25 for both selections, while TSCPS1_pos_0.25 indicates TSCPS 1 with positively coordinated samples and $\alpha =$ 0.25 for both selections).

Similarly, Figures 5.3 and 5.4 show boxplots corresponding to the $\hat{A}$ values obtained through Monte Carlo simulation, for both settings, respectively. The white boxplots correspond to the $\hat{A}$ values obtained from independent samples s_1 and s_2, while the grey ones to negatively coordinated samples s_1 and s_2. In all figures, LPM with PRNs as well SCPS with PRNs show smaller spread of the $\hat{D}$ values and $\hat{A}$ values compared to Poisson sampling designs since both provide fixed sample sizes and are able to manage the spatial correlation of the data.

Figures 5.1 and 5.2 show a similar pattern of the boxplots: a larger overlap between s_1 and s_2 leads to a smaller spread of the $\hat{D}$ values. As expected, the spread of the $\hat{D}$ values is reduced for each type of positively coordinated samples compared to independent samples' selection. For LPM and SCPS designs this reduction is, however, less important. This fact can be explained by the smaller overlap between positively coordinated samples in LPM and SCPS designs compared to the other ones, as the examples in Section 5.1 show it. The larger sample sizes in the second setting reduce the spread of the $\hat{D}$ values in the case of positively coordinated samples (grey boxplots) compared to the independent sample selection (white boxplots). In Figures 5.3 and 5.4, negative coordination reduces in general the spread of the $\hat{A}$ values compared to independent sample selection. As in Figures 5.1 and 5.2, this reduction is less important for LPM and SCPS compared for example to Poisson sampling and TSCPS 2.

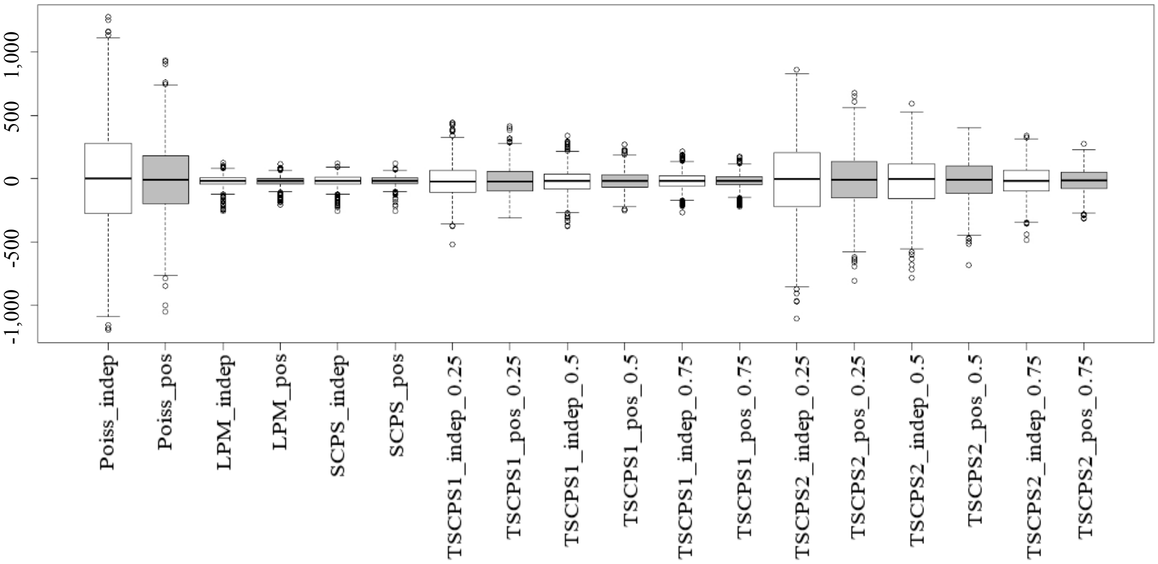


Figure 5.1 First setting: $N = 100, n_1 = 10, n_2 = 25,$ boxplots of the $\hat{D}$ values obtained through Monte Carlo simulation, the sampling design is specified below each boxplot. The white boxplots correspond to independent samples' selection, while the grey ones to positively coordinated samples.

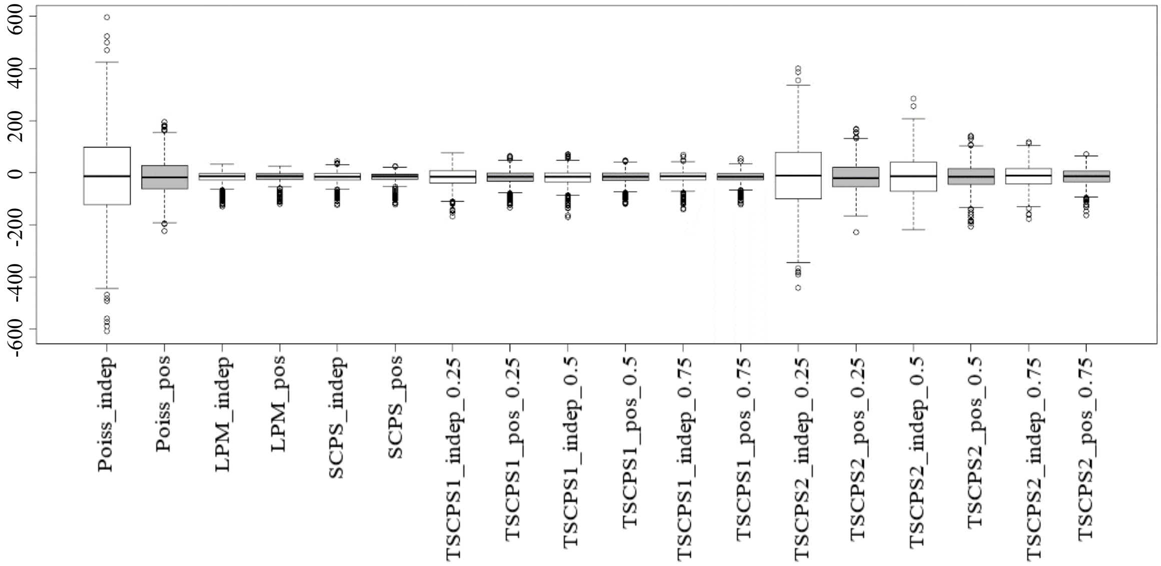


Figure 5.2 Second setting: $N = 100, n_1 = 50, n_2 = 50,$ boxplots of the $\hat{D}$ values obtained through Monte Carlo simulation, the sampling design is specified below each boxplot. The white boxplots correspond to independent samples' selection, while the grey ones to positively coordinated samples.

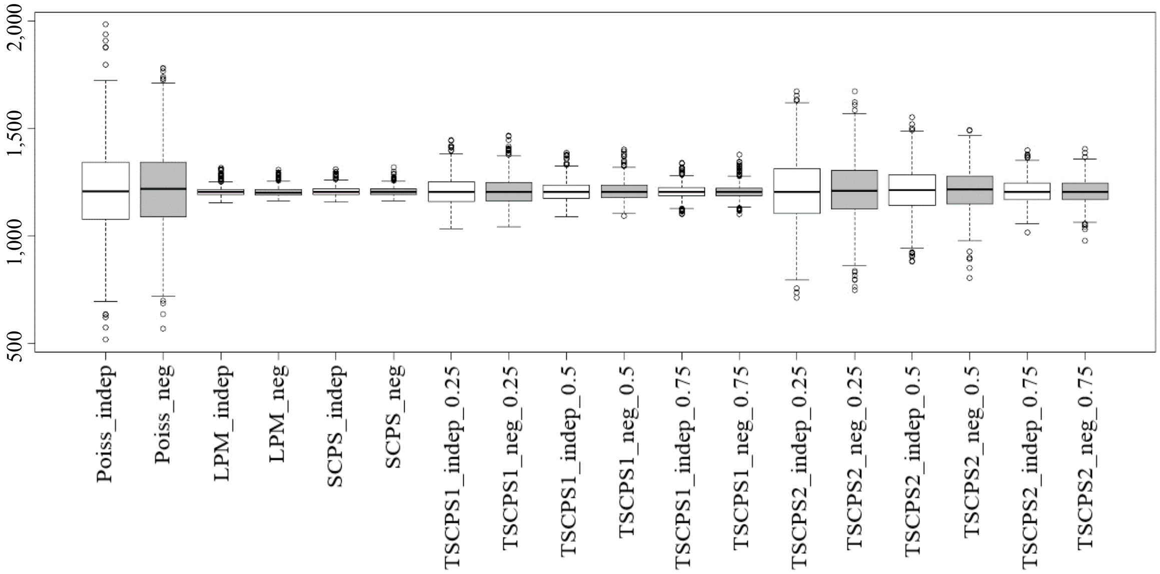


Figure 5.3 First setting: $N = 100, n_1 = 10, n_2 = 25,$ boxplots of the $\hat{A}$ values obtained through Monte Carlo simulation, the sampling design is specified below each boxplot. The white boxplots correspond to independent samples' selection, while the grey ones to negatively coordinated samples.

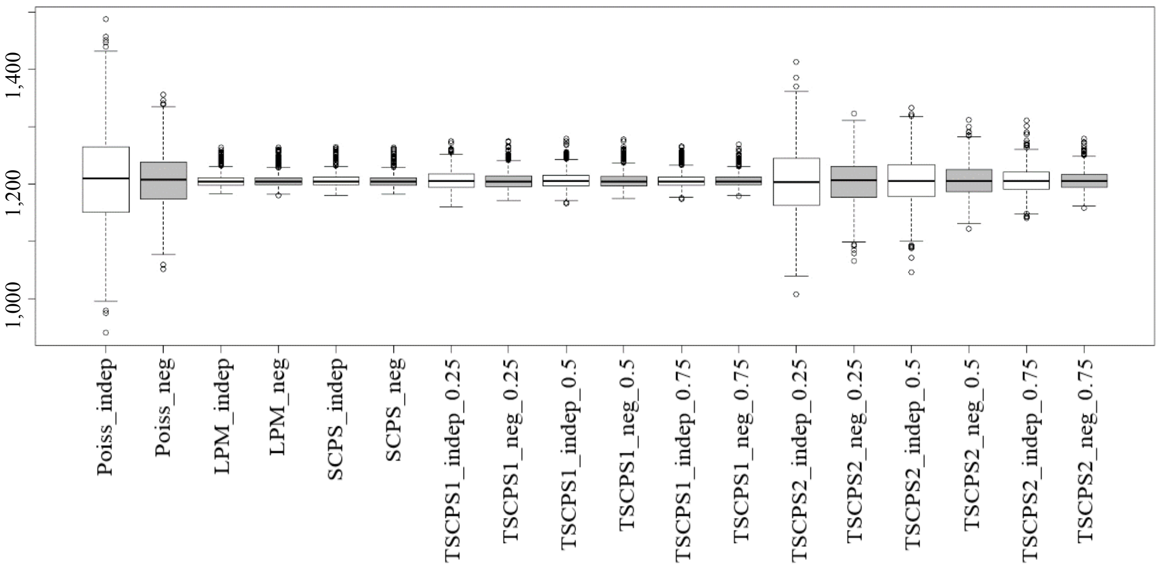


Figure 5.4 Second setting: $N = 100, n_1 = 50, n_2 = 50,$ boxplots of the $\hat{A}$ values obtained through Monte Carlo simulation, the sampling design is specified below each boxplot. The white boxplots correspond to independent samples' selection, while the grey ones to negatively coordinated samples.

To quantify the performance of the proposed methods, for positive and negative coordination, respectively, the Monte Carlo variance was used

$$\text{Var}_{\text{MC}}(\theta) = \frac{1}{m-1}\sum_{\ell=1}^{m}(\theta_\ell - E_{\text{sim}}(\theta))^2,$$

where θ_ℓ is the value of $\hat{D}$ or $\hat{A}$ obtained in the ℓ^{th} run and $E_{\text{sim}}(\theta) = \frac{1}{m}\sum_{j=1}^{m}\theta_j$. The reduction in variance estimation through overlapped samples of $\hat{D}$ is summarized in Table 5.9. The table shows the values of the ratio between $\text{var}_{\text{MC}}(\hat{D})$ obtained using positively coordinated samples and $\text{var}_{\text{MC}}(\hat{D})$ using independent samples for both settings. We note that for all sampling designs this ratio is less than 1, indicating a variance reduction through sample overlap. Table 5.10 shows the values of the ratio between $\text{var}_{\text{MC}}(\hat{A})$ obtained using negatively coordinated samples and $\text{var}_{\text{MC}}(\hat{A})$ using independent samples for both settings. For the first setting, except for Poisson sampling, the ratio is close to 1, showing negligible improvement of the negatively coordinated samples compared to independent selections. Using larger sample sizes, the second setting shows an important improvement for TSCPS 2, but not for LPM and SCPS.

Table 5.9
Ratio between $\text{var}_{\text{MC}}(\hat{D})$ obtained using positively coordinate samples and $\text{var}_{\text{MC}}(\hat{D})$ using independent samples

Design		$n_1 = 10, n_2 = 25$	$n_1 = 50, n_2 = 50$
		Ratio	Ratio
Poisson		0.481	0.178
LPM		0.759	0.679
SCPS		0.760	0.778
TSCPS 1	$\alpha = 0.25$	0.695	0.545
	$\alpha = 0.50$	0.739	0.700
	$\alpha = 0.75$	0.806	0.752
TSCPS 2	$\alpha = 0.25$	0.513	0.217
	$\alpha = 0.50$	0.571	0.319
	$\alpha = 0.75$	0.634	0.491

Table 5.10
Ratio between $\text{var}_{\text{MC}}(\hat{A})$ obtained using negatively coordinate samples and $\text{var}_{\text{MC}}(\hat{A})$ using independent samples

Design		$n_1 = 10, n_2 = 25$	$n_1 = 50, n_2 = 50$
		Ratio	Ratio
Poisson		0.792	0.324
LPM		0.941	0.949
SCPS		0.921	0.901
TSCPS 1	$\alpha = 0.25$	0.932	0.679
	$\alpha = 0.50$	0.950	0.840
	$\alpha = 0.75$	0.953	0.876
TSCPS 2	$\alpha = 0.25$	0.828	0.387
	$\alpha = 0.50$	0.834	0.463
	$\alpha = 0.75$	0.919	0.597

In summary, LPM with PRNs, SCPS with PRNs and the TSCPS family reduce the Monte-Carlo variance of the differences through sample overlap compared to independent samples' selection in both settings. For the independent samples' selection, these methods are more precise than Poisson sampling because they are able to manage the spatial trend present in the variable of interest, and the sample sizes are fixed (for LPM and SCPS using the "maximal weight strategy") or less variable than for Poisson sampling. The Monte-Carlo variance of the averages is negligibly reduced by LPM and SCPS using negatively coordinated samples compared to independent samples in both settings. The transformed SCPS family shows a real improvement in the second setting, when n_1 and n_2 are relatively large, for all α.

6 Application to Swiss establishments

We illustrate the application of the proposed methods on real data. The data that we used was collected by the Swiss Federal Statistical Office and can be downloaded for free (https://www.bfs.admin.ch/bfs/fr/home/services/geostat/geodonnees-statistique-federale/etablissements-emplois/statistique-structurel-entreprises-statent-depuis-2011.assetdetail.3303058.html). It contains census data from 2013 and 2015 on Swiss establishments. Data for all establishments are aggregated at the hectare level. The geographical coordinates are proper to each hectare, and not to establishments. Each hectare can contain several establishments. The statistical unit was in this application an hectare, and not an establishment. We considered only hectares containing establishments from the economic activity 1 (agriculture, hunting, forestry, fisheries and aquaculture), and having in total at least 3 full-time equivalent employees. The years 2013 and 2015 were considered the two time occasions. In 2013, a number of 7,057 units were available, while in 2015 this number was 7,104. The overall population was of size $N =$ 9,478. The difference in the sizes between the two time occasions was due to the 2,374 deaths and 2,421 births in 2015 compared to 2013. Figure 6.1 shows the geographical location of the units from the overall population. The parts inside of the figure with less locations correspond in majority to the Swiss Alps.

The data can be used with two main purposes:

- The location of each establishment in Switzerland has been geocoded since 1995. The register of establishments contains their geographical coordinates. Surveys are made to complete some missing information in this register. To achieve this, the Swiss Federal Statistical Office conducted such a survey in 2014. A positive coordination can be applied for example to check the quality of the the completed information from a time occasion to another one.
- Negative coordination can be applied to reduce the response burden of the establishments selected in several surveys. If the aggregated data are used, the hectares can be seen as primary selected units, while the establishments inside them as secondary units.

We used the values of the expected sample sizes $n_1 =$ 1,000 and $n_2 =$ 800, while $\pi_{i,1}$ and $\pi_{i,2}$ were computed proportional to the same variable measured in 2013 and 2015, respectively. This variables was the total number of full-time equivalent employees of all establishments inside of a hectar. A matrix of size $N \times N$ of PRNs was generated for the LPM. For the other methods, the vector of PRNs was taken to be the main diagonal of this matrix. In both time occasions respectively, we selected samples s_1 and s_2 using Poisson sampling with PRNs, LPM with PRNs, SCPS with PRNs, TSCPS 1 with PRNs (α = 0.25, 0.50, 0.75), and TSCPS2 with PRNs (α = 0.25, 0.50, 0.75). The Euclidean distance between locations was used in all methods, excepting Poisson sampling.

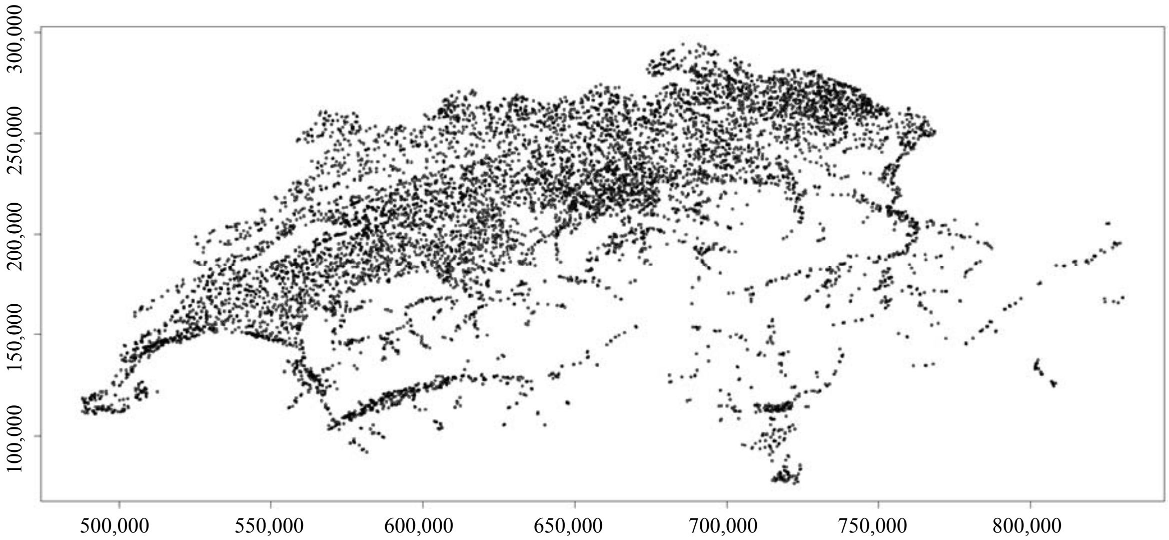


Figure 6.1 Swiss establishments aggregated data. Spatial distribution of the units in the overall population based on the census in 2013 and 2015.

We analyzed the selected samples in terms of realised overlap and B measure. To achieve this, positive and negative coordinations with PRNs were respectively applied. Table 6.1 shows the realised sample sizes as well as the overlap between different samples in both types of coordination. For the samples drawn in the first time occasion, the B measure given in expression (3.3) is also indicated. Poisson sampling presents the highest overlap in positive coordination (560, when AUB = 538.022), while LPM the smallest one. Due to the important changes in the population from 2013 to 2015, SCPS performs better than LPM, with an overlap of 329, but worse than Poisson sampling. All the members of the TSCPS family perform intermediately between Poisson sampling and SCPS, in function of the value of α. Negative coordination shows the same superiority of Poisson sampling, while the other designs exhibit smaller values of the realised overlap, with SCPS performing again better than LPM. Moving now to the spatial balancing feature, Poisson sampling yields the largest realised B measure, while LPM and SCPS as expected indicate the smallest ones. As in the results shown in Section 5.2, the members of the TSCPS family exhibit smaller realised B measure than Poisson sampling, but larger than SCPS. The application of the proposed methods on these real data indicates similar behavior of them with the simulation results shown in Sections 5.1 and 5.2.

Table 6.1
Swiss establishments aggregated data. N = 9,478, n_1 = 1,000, n_2 = 800, AUB = 538.022, ALB = 45.908. Realised sample sizes, overlap between s_1 and s_2 in both types of coordination, and the B measure for s_1

Design		size of s_1	Positive coord.		Negative coord.		B_{s_1}
			size of s_2	overlap	size of s_2	overlap	
Poisson		1,010	840	560	779	46	0.387
LPM		1,000	800	270	800	93	0.161
SCPS		1,000	800	329	800	70	0.151
TSCPS 1	$\alpha =$ 0.25	999	799	459	800	64	0.178
	$\alpha =$ 0.50	1,000	799	420	800	66	0.217
	$\alpha =$ 0.75	1,000	800	366	800	67	0.178
TSCPS 2	$\alpha =$ 0.25	1,012	830	469	808	49	0.275
	$\alpha =$ 0.50	1,020	828	409	799	58	0.194
	$\alpha =$ 0.75	1,010	816	377	797	66	0.153

7 Conclusions

New methods are proposed to coordinate spatially balanced samples based on PRNs. The objective is two-fold: first, to achieve a good coordination degree between samples, and second to preserve a good spatial balance degree. With the coordination of LPM and SCPS a good degree of spatial balance is ensured. SCPS with PRNs is less memory consuming since only a PRN vector of size N is used, while for LPM one uses a matrix of dimension $N \times N$. Our examples concern moderate size populations, and a large N quickly introduces limits in the calculations. In practice, a large N leads to an oversized matrix to be employed in the coordination of LPM samples. In these cases, the method can be implemented using dynamic allocation of the computer memory. Despite this solution, limits of the proposed method are possible in practice.

In our simulations, SCPS tends to perform better than LPM in terms of overlap expectation and variance, for both positive and negative coordination. A good coordination of LPM samples is more difficult to achieve than of SCPS samples, because the same pairs of units should be considered in the sample selection process, instead of single units. If births or deaths appear in the population, the pairs used for the selection of s_1 may not be available any more for the selection of s_2. Thus, the sample coordination becomes poor. SCPS does not have this weakness, but instead the coordination level may drop as well if the weights are distributed very differently the second time compared to the first time. This is the reason why SCPS with PRNs performs worse than Poisson sampling with PRNs. LPM with PRNs may have better behavior in terms of overlap than SCPS with PRNs if changes in the population are not detected. This situation is exemplified in Table 5.1 when the static MU284 population is used.

As shown in our examples in Section 5.1 both methods show a weaker performance in terms of expected overlap than Poisson sampling. This is a normal feature of these methods since one imposes the fixed sample size constraint for LPM and SCPS. In order to overcome this weakness, we introduced a new family of designs, based on a transformation of SCPS and a choice of scalar α, $0 \le \alpha \le 1$. Each value of α leads to a member of this family. For $\alpha = 1$ one obtains SCPS, while for $\alpha = 0$ Poisson sampling. This family of designs reminds us another family depending upon a scalar, the Pomix design (Kröger et al., 1999). The Pomix design is a mixture between Bernoulli and Poisson sampling, also used for coordination with PRNs.

For the transformed version of SCPS, the degree of coordination and spatial balance depend on the choice of α. Being a mixture of Poisson sampling and SCPS, it achieves a better coordination degree than SCPS. However, the improved degree of coordination comes at the cost of increased variance of sample size and reduced spatial balance as our examples in Section 5.1 and Section 4 showed. Based on our results, for the transformed SCPS, our recommendation is to use $\alpha = 0.5$ that represents a compromise between a good spatial balance degree and a good coordination degree. On the other hand, $\alpha = 0.5$ seems a good all-purpose suggestion since the results for variance estimation of differences and averages shown in Section 5.3 also indicate this value as a reasonable choice.

In our results shown in Section 5.3 LPM with PRNs, SCPS with PRNs and the TSCPS family reduce the Monte-Carlo variance of the differences when positively coordinated samples are used compared to independent samples' selection. In both used settings, it seems, however, that in the case of LPM with PRNs and SCPS with PRNs, variance reduction comes mainly from the combined effect of spatial balance and fixed sample size rather than from the effect of positive coordination. The Monte-Carlo variance of the averages is not always reduced in our examples when negatively coordinated samples are selected compared to independent samples; LPM with PRNs and SCPS with PRNs show in this case negligible improvement when negative coordination is used.

All the proposed methods can also be applied in the case where the spatial distance is replaced by a distance between auxiliary variables like the Mahalanobis distance. Thus, the sample coordination can be performed in the space spanned by these variables. The proposed methods allow thus not only a spatial sample coordination, but also the coordination of representative samples, in the terminology used by Grafström and Schelin (2014).

Aknowledgements

The authors wish to thank the Associate Editor and two referees for their valuable comments and suggestions that helped in improving the quality of the paper significantly.

References

Benedetti, R., Piersimoni, F. and Postiglione, P. (2017). Spatially balanced sampling: A review and a reappraisal. *International Statistical Review*, 85, 439-454.

Bondesson, L., and Thorburn, D. (2008). A list sequential sampling method suitable for real-time sampling. *Scandinavian Journal of Statistics*, 35, 466-483.

Brewer, K., Early, L. and Joyce, S. (1972). Selecting several samples from a single population. *Australian Journal of Statistics*, 3, 231-239.

Cotton, F., and Hesse, C. (1992). Tirages coordonnés d'échantillons. Technical Report E9206, Direction des Statistiques Économiques, INSEE, Paris, France.

Deville, J.-C., and Tillé, Y. (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika*, 85, 89-101.

Dickson, M.M., Benedetti, R., Giuliani, D. and Espa, G. (2014). The use of spatial sampling designs in business surveys. *Open Journal of Statistics*, 04, 345-354.

Dubin, R.A. (1992). Spatial autocorrelation and neighborhood quality. *Regional Science and Urban Economics*, 22, 3, 433-452.

GeoDa Center for Geospatial Analysis and Computation (2017). Sample data. http://spatial.uchicago.edu/sample-data. Accessed: 6-April-2017.

Grafström, A. (2012). Spatially correlated Poisson sampling. *Journal of Statistical Planning and Inference*, 142, 139-147.

Grafström, A., and Matei, A. (2015). Coordination of Conditional Poisson samples. *Journal of Official Statistics*, 31, 4, 649-672.

Grafström, A., and Schelin, L. (2014). How to select representative samples. *Scandinavian Journal of Statistics*, 41, 2, 277-290.

Grafström, A., and Tillé, Y. (2013). Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. *Environmetrics*, 14, 2, 120-131.

Grafström, A., Lundström, N.L.P. and Schelin, L. (2012). Spatially balanced sampling through the pivotal method. *Biometrics*, 68, 2, 514-520.

Haziza, D. (2013). Sampling and estimation procedures in business surveys: A discussion of some specific features. Seminar of the Royal Statistical Society, London, England.

Kröger, H., Särndal, C.-E. and Teikari, I. (1999). Poisson mixture sampling: A family of designs for coordinated selection using permanent random numbers. *Survey Methodology*, 25, 1, 3-11. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/1999001/article/4707-eng.pdf.

Kröger, H., Särndal, C.-E. and Teikari, I. (2003). Poisson mixture sampling combined with order sampling. *Journal of Official Statistics*, 19, 59-70.

Mach, L., Reiss, P.T. and Şchiopu-Kratina, I. (2006). Optimizing the expected overlap of survey samples via the northwest corner rule. *Journal of the American Statistical Association*, 101, 476, 1671-1679.

Matei, A., and Tillé, Y. (2005). Maximal and minimal sample co-ordination. *Sankhyā*, 67, part 3, 590-612.

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

Stevens, D.L.J., and Olsen, A.R. (2004). Spatially balanced sampling of natural resources. *Journal of the American Statistical Association*, 99, 262-278.

Survey Methodology, December 2018
Vol. 44, No. 2, pp. 239-252
Statistics Canada, Catalogue No. 12-001-X

Using balanced sampling in creel surveys

Ibrahima Ousmane Ida, Louis-Paul Rivest and Gaétan Daigle[1]

Abstract

These last years, balanced sampling techniques have experienced a recrudescence of interest. They constrain the Horvitz Thompson estimators of the totals of auxiliary variables to be equal, at least approximately, to the corresponding true totals, to avoid the occurrence of bad samples. Several procedures are available to carry out balanced sampling; there is the cube method, see Deville and Tillé (2004), and an alternative, the rejective algorithm introduced by Hájek (1964). After a brief review of these sampling methods, motivated by the planning of an angler survey, we investigate using Monte Carlo simulations, the survey designs produced by these two sampling algorithms.

Key Words: Balanced sampling; Creel surveys; Cube method; Multistage sampling; Rejective algorithm; Monte Carlo simulation.

1 Introduction

Creel surveys provide the foundation for estimating the impact of recreational fishing (Pollock, Jones and Brown, 1994). They are conducted to estimate total catch, fishing effort, and catch rate for various species at several locations (Hoenig, Jones, Pollock, Robson and Wade, 1997). As they focus on fish of interest to recreational anglers, they provide useful information for the management and economic contribution of sport fisheries (Minnesota Department of Natural Resources, 2011).

Two methods are used to contact anglers in creel surveys, either the site access or the roving method. In site access, an agent waits at a location that the anglers must go through when they leave the site and interviews them when they depart (Robson and Jones, 1989). With the roving method the agent moves through the survey area and contacts anglers while they are fishing (United States Environmental Protection Agency, 1998). As the agent cannot be on location for the whole survey, survey sampling is used to select the periods when he will be on site, interviewing fishermen.

In practice creel surveys can face several operational constraints especially when they involve many sites as an agent can only be at one site at a given time. Accommodating all these constraints can be a real challenge when planning a survey. This paper discusses balanced sampling in this context. By framing some operational constraints as balancing equations in a multi-stage sampling design, one should be able to ensure that the sample selected meets the necessary requirements.

Balanced sampling is reviewed in Tillé (2011). A popular method to select a balanced sample is the cube method of Deville and Tillé (2004). An alternative is to select repeatedly several unbalanced samples until, by chance, a sample that approximately meets the balancing equations is drawn. This is the rejective method introduced by Hájek (1964), see also Fuller (2009) and Legg and Yu (2010). In a creel survey, the number of balancing equations is typically large. The implementation of the cube method in this context is discussed

[1] Ibrahima Ousmane Ida, Louis-Paul Rivest and Gaétan Daigle, Université Laval. E-mail: louis-paul.rivest@mat.ulaval.ca.

in Chauvet (2009) and Hasler and Tillé (2014). See Vallée, Ferland-Raymond, Rivest and Tillé (2015) for a recent application of these methods in the context of a forest inventory. A recent paper in this area by Chauvet, Haziza and Lesage (2015) investigates the properties of the balanced samples obtained using a rejective method.

The objectives of this paper are twofold. First, the operational constraints for a creel survey of striped bass (*Morone saxatilis*) carried out in the Gaspé Peninsula are presented. Then we will show how balanced sampling, implemented using the cube method, can be used to plan a survey fulfilling most of the constraints. The last section of the paper compares the rejective method to the cube method in the context of creel surveys.

In Section 2, balanced sampling is presented using either the cube method or rejective sampling. Section 3 introduces operational constraints for a creel survey and shows how they can be met using balanced sampling with the cube method. In Section 4, the cube method is compared with the rejective algorithm in the context of a resource inventory where the balancing equations only involve indicator variables. Discussions of the results are presented in the Section 5.

2 Balanced sampling

Suppose that U is a finite population of size N that is sampled with a design having selection probabilities given by $\{\pi_i\!: i = 1, \ldots, N\}.$ If x is an auxiliary variable known for all population units, then the sample is balanced on x if the Horvitz-Thompson estimator for the total of x is equal to the known total of $x.$ In other words, for any balanced sample $s,$ the following equation has to be satisfied,

$$\sum_{i \in s} \frac{x_i}{\pi_i} = \sum_{i=1}^{N} x_i. \tag{2.1}$$

For the surveys considered here, we balance on indicator variables $I_i\,(\omega)$ equal to 1 if unit i is of type ω and 0 otherwise. If all the units i for which $I_i\,(\omega)$ is equal to 1 have the same selection probability $\pi_\omega,$ then equation (2.1) reduces to $\sum_{i \in s} I_i\,(\omega) \big/ \pi_\omega = \sum_{i=1}^{N} I_i\,(\omega).$ In this context the balancing equation simply requests that the number of sampled units of type $\omega,\, n_\omega = \sum_{i \in s} I_i\,(\omega)$, is equal to its expectation,

$$n_\omega = \sum_{i=1}^{N} I_i\,(\omega)\,\pi_\omega. \tag{2.2}$$

To implement balanced sampling we use the cube method of Deville and Tillé (2004), and the extension of Hasler and Tillé (2014) to cope with highly stratified populations. In Section 4 this method is compared with the implementation of the rejection method proposed by Fuller (2009). In the context of this study, we are balancing on T types of units; we want the sampled numbers of units for the T types, $\tilde{n} = (n_1, \ldots, n_T)^\top,$ to be equal to their expectations, $E\,(\tilde{n})$, under the sampling design. Under rejective sampling, the sample is said to be balanced if

$$Q_{T,\,n} = (\tilde{n} - E\,(\tilde{n}))^\top\,[\mathrm{Var}\,(\tilde{n})]^{-1}\,(\tilde{n} - E\,(\tilde{n})) < \gamma^2 \tag{2.3}$$

where $\operatorname{Var}(\tilde{n})$ represents the design based covariance matrix of $\tilde{n}$ and γ^2 is a tolerance value that determines the balancing condition. Samples that do not meet the balancing equation $Q_{T,n} < \gamma^2$ are simply rejected.

3 A creel survey for striped bass in the Gaspé Peninsula

The Gaspé Peninsula is on the Canadian East Coast in the Province of Québec. In 2015 a creel survey for striped bass was conducted in this peninsula as recreational striped bass fishing had just been reintroduced after a long moratorium.

The study area, presented in Figure 3.1, is scattered over more than 250 kms, on the Gaspé Peninsula coast. The survey is carried out by a single wildlife agent; it is not possible for him to visit two distant sites on the same day. For that reason, neighboring sites are grouped into three sectors as shown in Figure 3.1. We consider the survey for the 33 holidays. The survey variable is the fishing effort, in number of hours of fishing. As some sites attract more fishermen than others, the number of visits to site l of sector i has to be proportional to its importance x_{il} as given in Table 3.1. In addition, for the purpose of the survey, a day is divided into three periods (AM, PM, EV), where EV stands for evening, and six subperiods (AM1, AM2, PM1, PM2, and EV1, EV2). For instance AM1 goes from 8:00 to 10:00 while AM2 is from 10:00 to 12:00. A working day contains two periods and four subperiods. For instance if the agent works AM and PM, then he has a free evening. Thus during a working day he is able to visit four sites, two per working period.

The survey population on a day consists of 54 quadruplets, $(\text{sector} \times \text{period} \times \text{subperiod} \times \text{site})$, 4 of which are sampled. To denote population units the following indices are useful:

i) $h = 1, \ldots, H = 33$ represents the days;

ii) $i = 1, 2, 3$ stands for the sectors in Figure 3.1;

iii) $j = 1, 2, 3$ denotes a period within a day;

iv) $k = 1, 2$ represents the subperiods within a period;

v) $l = 1, 2, 3$ represents the sites, see Figure 3.1, within a sector.

The goal is to estimate the fishing effort for combination of subperiod (6 levels) and site (9 levels). We want to plan a survey with a predetermined sample size for the 54 cells of the cross-classified table. The basic selection probabilities are

$$\pi_{hijkl} = \frac{2x_{il}}{3x_{\bullet\bullet}}, \tag{3.1}$$

where replacing i or l by $\bullet$ means that a summation is taken on the corresponding index. Observe that the sum of π_{hijkl} over the indices (i, j, k, l) is equal to 4, the number of units visited by the wildlife technician on a single day.

At a first glance, the sample could possibly be drawn in a single stage using selection probabilities (3.1) by balancing on the 54 site by subperiod indicator variables. This is not feasible because of operational constraints. The first one is that on a single day the technician visits sites from the same sector to limit the traveling between sites. The second constraint is that on a working day the technician is off duty for the two subperiods of the same period. In order to meet these operational constraints we propose, in the next section, a design having three levels of sampling where sectors are selected at level 1, periods are selected at level 2 and sites are selected at level 3.

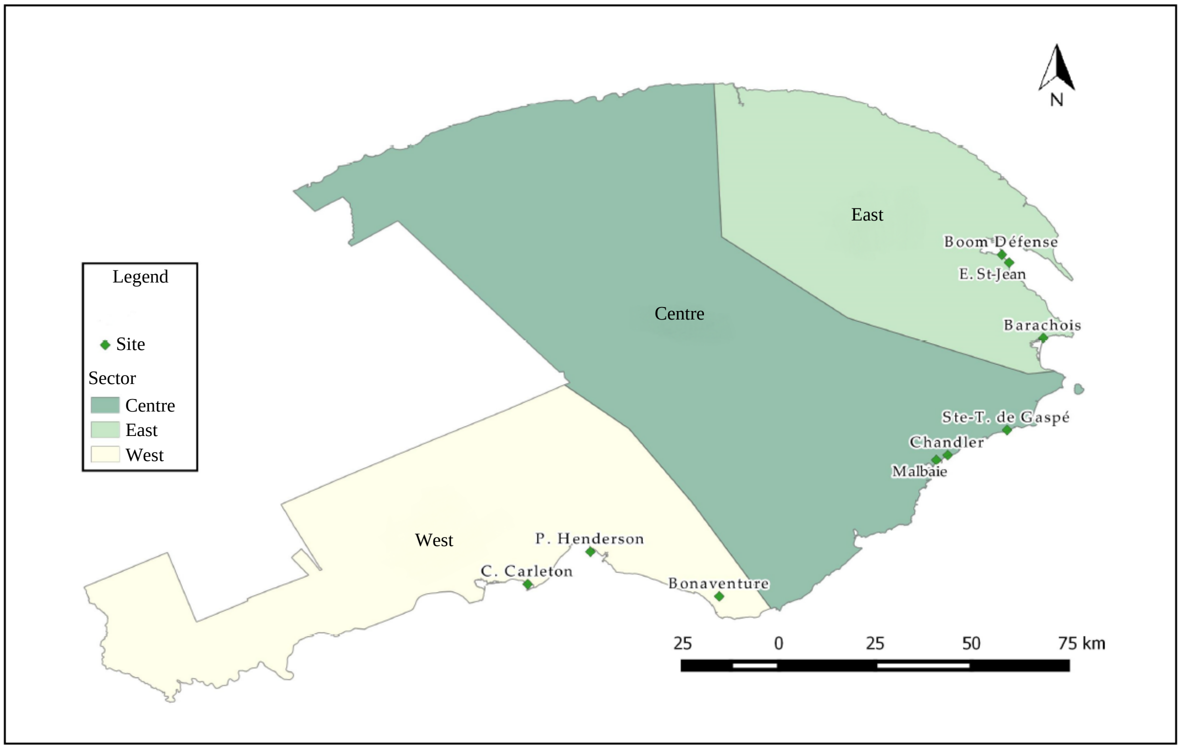


Figure 3.1 The 9 sites to be surveyed for striped bass.

Table 3.1
Average and expected number of visits to each site

Sector	Site	x_{il}	$E\left(n_{il}\right)$	$\bar{n}_{il}$	$\mathrm{Sd}_{n_{il}}$
East $(i = 1)$	Boom Défense $(l = 1)$	2	20.308	20.286	0.850
	E. St-Jean $(l = 2)$	1	10.154	10.153	0.621
	Barachois $(l = 3)$	2	20.308	20.296	0.881
Centre $(i = 2)$	Ste-T. de Gaspé $(l = 4)$	1	10.154	10.176	0.865
	Malbaie $(l = 5)$	1	10.154	10.155	0.880
	Chandler $(l = 6)$	1	10.154	10.162	0.881
West $(i = 3)$	Bonaventure $(l = 7)$	2	20.308	20.311	1.004
	P. Henderson $(l = 8)$	1	10.154	10.153	0.681
	C. Carleton $(l = 9)$	2	20.308	20.309	1.016

3.1 A balanced multi-stage design for creel survey

This section describes the three stages of the survey that ensures that the operational constraints presented in the previous section are met. It also gives, for each stage, the balancing variables.

The first stage is stratified by day; for each day a single sector is drawn with selection probabilities $x_{i\bullet} / x_{\bullet\bullet}$. At level two, for each sector selected at level 1, two periods are selected out of 3 using simple random sampling (i.e., with selection probabilities 2/3). At level three, a sector*period is stratified by subperiod and one site is selected for each subperiod, the selection probabilities are $x_{il} / x_{i\bullet}$. In summary the selection probabilities at the three levels are

$$\pi_{hi}^{(1)} = \frac{x_{i\bullet}}{x_{\bullet\bullet}}, \qquad \pi_{j|i}^{(2)} = \frac{2}{3}, \qquad \pi_{l|ijk}^{(3)} = \frac{x_{il}}{x_{i\bullet}}.$$

As expected the product $\pi_{hi}^{(1)} \times \pi_{j|i}^{(2)} \times \pi_{l|ijk}^{(3)}$ is equal to (3.1), the target selection probability.

The goal is still to get a sample with predetermined sample sizes for the 54 site by subperiod combinations. Thus balanced sampling needs to be implemented at each stage. At level 1 we need to balance on the indicator variables for the three sectors while at level 2 balancing on the 9 indicator variables for the sector by period combinations is needed. Balancing at level 3 is slightly more complicated as it involves several strata.

At level 2, $33 \times 2 = 66$ sector*periods have been selected. Each one is stratified by subperiod so we are facing 132 strata at level 3 and one site is selected from each one. Balancing is needed with respect to the 54 site by subperiod indicator functions. This is a complex problem and the balancing constraints (2.3) involve the inverse of a large variance covariance matrix. Thus to implement a rejective algorithm in this context one would need an alternative to criterion (2.3) for accepting a sample. For now we discuss the implementation of balanced sampling for this design with the cube method. Comparisons between the cube method and rejective sampling in the context of a simplified creel survey are presented in Section 4.

Among the 132 third stage strata, the number of strata for one subperiod, say AM2, in sector i is an integer close to $22x_{i\bullet} / x_{\bullet\bullet}$ that depends on the stage 2 sample. This integer plays the role of $\sum_{i=1}^{N} I_i\ (\omega)$ in equation (2.2) for balancing the sites of sector i at stage 3 while, for the l^{th} site, the probability in (2.2) is $\pi_\omega = x_{il} / x_{i\bullet}$. The stage 3 calibration equations for the 54 site by subperiod indicator functions can be described in a similar way. Clearly, it is not possible to meet exactly the 54 balancing equations and the cube method will give a sample that is approximately balanced.

The approximation occurs at the landing phase of the algorithm where balancing constraints are dropped in order to complete the selection of the sample, as introduced in Deville and Tillé (2004). As the stage 3 sample is highly stratified, we use the implementation of the landing phase in the function `balancedstratification2` developed in Hasler and Tillé (2014), with a small correction that prevents it from stopping when the sample is already balanced at the start of the landing phase. In the matrix of balancing constraints, the site constraints were given more importance than those which make visits to each site equally distributed among subperiods at level 3. They were the last ones to be dropped at the landing phase of the cube method.

To investigate how a failure to meet all balancing equations impacted the sample design, we generated $B = 10{,}000$ random replications of the balanced sample. The number of visits n_{il} to site (i, l) was noted. Table 3.1 compares the average $\overline{n}_{il}$ of n_{il} over the Monte Carlo replications,

$$\overline{n}_{il} = \frac{1}{B} \sum_{b=1}^{B} n_{il}^{(r)},$$

to its expectation, $E\,(n_{il})$. For all practical purposes, the two are equal and a failure to meet some balancing equations has no impact on the site selection probabilities. Table 3.1 also reports the standard deviations

$$\mathrm{Sd}_{n_{il}} = \left\{ \frac{1}{B-1} \sum_{b=1}^{B} \left(n_{il}^{(b)} - \overline{n}_{il} \right)^2 \right\}^{1/2}. \tag{3.2}$$

Most of the standard deviations are less than 1 in Table 3.1. Thus the absolute differences between target and realized sample sizes are less than or equal to 2 for most Monte Carlo samples.

Table 3.2 gives the expected number of visits in the 6 subperiods; they are all equal to 22, up to two decimal points, with standard deviations less than 0.2. Thus the period and subperiod constraints are met. Table 3.3 gives a realized sample for the first five days of the creel survey. It shows a harmonious permutation of sectors at level 1, periods at level 2, and sites at level 3 through the days because of the way in which the sample design was constructed. Given a balanced sample produced by the cube algorithm, an arbitrary permutation of the days gives an alternative balanced sample. Indeed the sampling design is invariant to a relabeling of the days. For instance, with the sample of Table 3.3 the technician has to travel from the western to the eastern sector between days 4 and 5. To avoid this long trip one could interchange days 1 and 5: the first two days would then be spent in the eastern sector and between days 4 and 5 the technician would travel from the western to the central sector. The alternative and the original samples have the same estimated totals for the calibration variables.

Table 3.2
Average and expected number of visits at each subperiod

Period	**Subperiod**	$E\left(n_{jk}\right)$	$\overline{n}_{jk}$	$\mathrm{Sd}_{n_{jk}}$
Morning ($j = 1$)	8h00-10h00 ($k = 1$)	22	22.000	0.000
	10h00-12h00 ($k = 2$)	22	22.000	0.000
Afternoon ($j = 2$)	12h00-15h00 ($k = 3$)	22	21.999	0.184
	15h00-18h00 ($k = 4$)	22	21.999	0.184
Evening ($j = 3$)	18h00-20h30 ($k = 5$)	22	22.001	0.184
	20h30-23h00 ($k = 6$)	22	22.001	0.184

Table 3.3
Units selected in a balanced sample for the first five days

H	Sector	Period	Subperiod	Site
1	Centre $(i = 2)$	Afternoon $(j = 2)$	12h00-15h00 $(k = 3)$	Chandler $(l = 6)$
			15h00-18h00 $(k = 4)$	Malbaie $(l = 5)$
		Evening $(j = 3)$	18h00-20h30 $(k = 5)$	Chandler $(l = 6)$
			20h30-23h00 $(k = 6)$	Ste-T. de Gaspé $(l = 4)$
2	East $(i = 1)$	Morning $(j = 1)$	8h00-10h00 $(k = 1)$	E. St-Jean $(l = 2)$
			10h00-12h00 $(k = 2)$	Boom Défense $(l = 1)$
		Evening $(j = 3)$	18h00-20h30 $(k = 5)$	Barachois $(l = 3)$
			20h30-23h00 $(k = 6)$	E. St-Jean $(l = 2)$
3	Centre $(i = 2)$	Morning $(j = 1)$	8h00-10h00 $(k = 1)$	Malbaie $(l = 5)$
			10h00-12h00 $(k = 2)$	Ste-T. de Gaspé $(l = 4)$
		Afternoon $(j = 2)$	12h00-15h00 $(k = 3)$	Malbaie $(l = 5)$
			15h00-18h00 $(k = 4)$	Chandler $(l = 6)$
4	West $(i = 3)$	Morning $(j = 1)$	8h00-10h00 $(k = 1)$	P. Henderson $(l = 8)$
			10h00-12h00 $(k = 2)$	Bonaventure $(l = 7)$
		Afternoon $(j = 2)$	12h00-15h00 $(k = 3)$	C. Carleton $(l = 9)$
			15h00-18h00 $(k = 4)$	C. Carleton $(l = 9)$
5	East $(i = 1)$	Afternoon $(j = 2)$	12h00-15h00 $(k = 3)$	Boom Défense $(l = 1)$
			15h00-18h00 $(k = 4)$	Barachois $(l = 3)$
		Evening $(j = 3)$	18h00-20h30 $(k = 5)$	Boom Défense $(l = 1)$
			20h30-23h00 $(k = 6)$	Barachois $(l = 3)$

3.2 Estimation of the fishing effort and of its variance

Once the survey is completed, the sample is a set of site × subperiod $\{(h, i, j, k, l)\}$ with sampling weights equal to the inverse of the selection probabilities given in (3.1). As the balancing equations for the 54 cells of the site by subperiod cross-classified table are not met exactly, we propose, following Deville and Tillé (2004), calibrating the survey weights on the total, H, of the indicator variables for these 54 cells. All the sampled units in cell (i, j, k, l) have the same weight, namely $1/\pi_{ijkl}$ where $\pi_{ijkl} = \pi_{hijkl}$, defined in (3.1), does not depend on h. The calibrated weight for a sampled unit in cell (i, j, k, l) is

$$w_{ijkl}^{(c)} = \frac{1}{\pi_{ijkl}} \times \frac{H}{n_{ijkl}/\pi_{ijkl}} = \frac{H}{n_{ijkl}},$$

where n_{ijkl} is the sample size for cell (i, j, k, l); it is the number of days for which site l of sector i has been visited during subperiod k of period j. In general n_{ijkl} is a random variable. When the samples are perfectly balanced, (2.2) implies that $n_{ijkl} = H\pi_{ijkl}$; the calibrated and basic weights are then equal. Now if y_{hijkl} represents the fishing effort for population unit (h, i, j, k, l), the fishing effort in cell (i, j, k, l) is $Y_{Uijkl} = \sum_h y_{hijkl}$. Its calibrated estimator is $\hat{Y}_{ijkl} = H\bar{y}_{sijkl}$ where $\bar{y}_{sijkl}$ is the average fishing effort for the

n_{ijkl} units sampled for that cell of the cross classified table. An estimator for the total fishing effort is obtained by summing the cells' estimated totals.

The evaluation of a design based variance estimator for the calibrated estimator of the total fishing effort is complex. A simple variance estimator for the estimated total for a single cell of the cross-classified table is available. The sample of days selected for cell (i, j, k, l) is a Bernoulli sample with selection probabilities π_{ijkl}, neglecting the balancing constraints. Thus by conditioning on the sample size, n_{ijkl}, $\hat{Y}_{ijkl}$ is H times the sample mean of a simple random sample. It is a design-unbiased estimator whose variance can be estimated using the formula for the variance of an estimated total in a simple random sampling design. We claim that these results are still valid when the balancing constraints are taken into account since the balanced sample design is invariant to a relabelling of the days. The estimated fishing efforts for the 54 cells of the cross-classified table are however dependent and it seems difficult to come up with a conditionally unbiased design based variance estimator for their total. A model based estimator seems to be only approach available for this total.

For the survey actually conducted in 2015, the methods used to estimate fishing effort and total catch are among those proposed in Pollock et al. (1994). It was a roving survey and the fishing effort at a sampled site was calculated as the average number of anglers on the site during the subperiod times the length, in hours, of the subperiod. Fishing efforts were estimated using calibrated weights; additional results are available in (Daigle, Crépeau, Bujold and Legault, 2015).

4 Comparison of the cube method and the rejective algorithm

Chauvet et al. (2015) have studied the cube method and the rejective algorithm by examining different aspects of these balancing techniques. They balanced on continuous auxiliary variables and they documented how the balancing algorithm impacted the selection probabilities and the sampling properties of estimators of population totals. The goal of this section is to compare the two sampling algorithms in a resource inventory where the balancing equations only involve indicator variables. This comparison is carried out in the context of a simplified creel survey with a stratified two stage design. The days represent strata $h = 1, \ldots, H,$ the sectors are defined as primary units $i = 1, 2, 3$ and sites, indexed by $j,$ are the secondary units. This sampling plan is similar to the design exposed in Section 3.1 except that periods and subperiods do not enter in the sampling design.

On each day two out of 3 sectors are selected and within each one 2 sites are sampled; thus 4 units are selected each day. The site importance variable x_{ij} determines the inclusion probabilities π_{hij} = $(2x_{i\bullet} / x_{\bullet\bullet}) \times (2x_{ij} / x_{i\bullet}) = \pi_{hi} \times \pi_{hj|i}$ for the two stages. As two out of three units are selected at each level, the joint selection probabilities are completely determined by $\{(\pi_{hi}, \pi_{hj|i}) : i,\ j = 1, 2, 3\}$ for the two stages; see the Appendix. If Z_{hij} stands for the indicator variables taking the value 1 if site $(i,\ j)$ is sampled on day h and 0 otherwise then the entries of 9×9 variance covariance matrix for $\{Z_{hij}: i,\ j = 1, 2, 3\}$ are given by

$$\operatorname{Cov}\left(Z_{hij}, Z_{hi'j'}\right) = \begin{cases} \pi_{hij} - \pi_{hij}^2 & \text{if } i = i' \text{ and } j = j' \\ \pi_{hi}\ \pi_{hjj'|i} - \pi_{hij}\ \pi_{hij'} & \text{if } i = i' \text{ and } j \neq j' \\ \pi_{hii'}\ \pi_{hj|i}\ \pi_{hj'|i'} - \pi_{hij}\ \pi_{hi'j'} & \text{if } i \neq i' \end{cases} \tag{4.1}$$

where $\pi_{hii'}$ represents the joint selection probability of sectors i and i' on a single day, $\pi_{hj|i}$ is the probability for selecting site $j,$ in sector $i,$ at stage 2 and $\pi_{hjj'|i}$ is the joint selection probability of sites j and j' in sector i. All these probabilities are evaluated using the size measure $x.$ Details are available in the appendix, see also Ousmane Ida (2016). The corresponding matrix $\operatorname{Var}(\tilde{n})$ in (2.3) is singular as one of the 9 constraints is redundant; thus in (2.3) a generalized inverse of the covariance matrix was used and $\gamma^2,$ in (2.3), was set equal to 2.73 and 7.34, the 5^{th} and the 50^{th} percentiles of the χ_8^2 distribution.

4.1 Simulations on the comparison of the cube method and of the rejective algorithm

To investigate the impact of the algorithm on the sampling properties of survey estimators we simulated, for each unit, a fishing effort for site (i, j) on day $h,$ $y_{hij},$ using independent Poisson random variables with mean $15 \times x_{ij}$. The total fishing effort for site (i, j) is then

$$Y_{Uij} = \sum_{h=1}^{H} y_{hij}.$$

A calibrated estimator, as defined in Section 3.2, for the fishing effort in site (i, j) is $\hat{Y}_{ij} = H\,\bar{y}_{sij},$ the average fishing effort for the n_{ij} units sampled at site (i, j) times H.

To compare the balancing algorithms, we used designs with $H = 12$ strata and two importance variables $x,$ one with a small variation between site and one with a medium variation. Under each scenario we generated $B = 100{,}000$ random replications of a balanced sample by using the cube methods on one hand, and two rejective algorithms on the other. The inclusion probabilities for site (i, j) was estimated by

$$\hat{\pi}_{ij} = \frac{1}{B \times H} \sum_{b=1}^{B} n_{ij}^{(b)}.$$

This estimator assumes that the inclusion probabilities π_{hij} are constant in $h.$ This holds true because the sample design is invariant to a relabelling of the days, see Section 3.1.

As argued in Section 3.2, the calibrated estimator $\hat{Y}_{ij}$ is design unbiased under the two selection algorithms. We compare their standard deviations,

$$\operatorname{Sd}_{\hat{Y}_{ij}} = \left\{ \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{Y}_{ij}^{(b)} - \bar{\hat{Y}}_{ij} \right)^2 \right\}^{1/2},$$

where $\bar{\hat{Y}}_{ij}$ is the average of the B simulated values. The sample size standard deviations were also calculated using (3.2). Observe that $\hat{\pi}_{ij} = \bar{n}_{ij} / H$. The simulation results are presented in Tables 4.1, 4.2 and 4.3.

Table 4.1
Comparison of the cube method (CM) and of two rejective algorithms (*R* 5% and *R* 50%) when *x* has a low variation

				CM		*R* 5%		*R* 50%	
Sector	Site	x_{ij}	π_{ij}	$\bar{\hat{\pi}}_{ij}$	$\mathrm{Sd}_{\hat{Y}_{ij}}$	$\bar{\hat{\pi}}_{ij}$	$\mathrm{Sd}_{\hat{Y}_{ij}}$	$\bar{\hat{\pi}}_{ij}$	$\mathrm{Sd}_{\hat{Y}_{ij}}$
$i = 1$	$j = 1$	3	0.500	0.500	16.56	0.503	16.86	0.505	17.40
	$j = 2$	2	0.333	0.333	22.20	0.329	23.35	0.328	25.07
	$j = 3$	3	0.500	0.500	23.99	0.503	24.47	0.505	25.15
$i = 2$	$j = 4$	2	0.333	0.333	25.80	0.329	26.93	0.326	29.11
	$j = 5$	2	0.333	0.333	33.97	0.329	35.54	0.326	38.28
	$j = 6$	2	0.333	0.333	27.65	0.329	28.87	0.326	31.10
$i = 3$	$j = 7$	3	0.500	0.500	22.50	0.502	22.88	0.502	23.66
	$j = 8$	3	0.500	0.500	20.02	0.502	20.20	0.502	20.94
	$j = 9$	4	0.667	0.667	22.01	0.674	21.98	0.679	22.25

Table 4.2
Comparison of the cube method (CM) and of two rejective algorithms (*R* 5% and *R* 50%) when *x* has a medium variation

				CM		*R* 5%		*R* 50%	
Sector	Site	x_{ij}	π_{ij}	$\bar{\hat{\pi}}_{ij}$	$\mathrm{Sd}_{\hat{Y}_{ij}}$	$\bar{\hat{\pi}}_{ij}$	$\mathrm{Sd}_{\hat{Y}_{ij}}$	$\bar{\hat{\pi}}_{ij}$	$\mathrm{Sd}_{\hat{Y}_{ij}}$
$i = 1$	$j = 1$	3	0.500	0.500	25.52	0.505	25.78	0.507	26.60
	$j = 2$	2	0.333	0.333	25.25	0.330	26.26	0.329	28.16
	$j = 3$	3	0.500	0.500	21.12	0.505	21.36	0.507	22.03
$i = 2$	$j = 4$	1	0.167	0.167	29.17	0.158	32.45	0.149	31.19
	$j = 5$	2	0.333	0.333	13.73	0.329	14.38	0.326	15.49
	$j = 6$	2	0.333	0.333	32.82	0.329	34.22	0.326	36.91
$i = 3$	$j = 7$	2	0.333	0.333	16.84	0.329	17.52	0.325	18.85
	$j = 8$	4	0.667	0.667	18.68	0.672	18.70	0.678	18.89
	$j = 9$	5	0.833	0.833	8.06	0.844	7.81	0.854	7.67

Table 4.3
Standard deviations of the sample sizes obtained with the cube method (CM) and with two rejective algorithms (*R* 5%, *R* 50%)

		x has a low variation				*x* has a medium variation			
Sector	Site	*x*	CM	*R* 5%	*R* 50%	*x*	CM	*R* 5%	*R* 50%
$i = 1$	$j = 1$	3	0.000	0.894	1.371	3	0.000	0.891	1.371
	$j = 2$	2	0.000	0.854	1.295	2	0.000	0.831	1.294
	$j = 3$	3	0.000	0.896	1.377	3	0.000	0.891	1.374
$i = 2$	$j = 4$	2	0.130	0.828	1.293	1	0.144	0.654	1.013
	$j = 5$	2	0.195	0.832	1.298	2	0.170	0.831	1.290
	$j = 6$	2	0.179	0.826	1.296	2	0.141	0.830	1.297
$i = 3$	$j = 7$	3	0.339	0.859	1.366	2	0.342	0.835	1.294
	$j = 8$	3	0.381	0.859	1.367	4	0.350	0.807	1.294
	$j = 9$	4	0.319	0.822	1.288	5	0.248	0.655	1.010

In Tables 4.1 and 4.2, the cube method maintains the selection probabilities and yields a total estimator with the smallest standard deviations. Taking γ^2 equal to the 50^{th} percentile of the χ_8^2 distribution for the rejective algorithm yields the poorer results, both in terms of selection probabilities and of the standard deviations of $\bar{y}_{sij}$. The largest biases for the selection probabilities occur at the extreme x values in Table 4.2. The selection probability for site $j = 4$ is underestimated by 11% with the rejective method based on the 50^{th} percentile and by 5% with the 5^{th} percentile. The probability is over estimated in the sites with the large values for x.

In Tables 4.1 and 4.2, the standard deviation for $\hat{Y}_{ij}$ is, in most cases, smallest for the cube method and largest for the rejection algorithm based on the 50^{th} percentile. The standard deviations for the rejective algorithm are up to 10% larger than the ones for the cube method. In Table 4.2, the largest gain in efficiency of the cube method with respect to the R 5% rejective algorithm (equal to the ratio of standard deviations squared) is 23%; it occurs when $j = 4$ and $x = 1$. These standard deviations are driven by the variability in sample sizes n_{ij}. Table 4.3 gives the sample sizes' standard deviations. Since the expected number of visits to sector 1 and to sites 1, 2, and 3 are integers, the cube method is able to get sample sizes equal to their expectations for this sector and the sample sizes standard deviations are 0. This is not possible in sectors 2 and 3 as the expected sample sizes for these sectors are not integer valued. In general, the rejective algorithms give sample sizes whose standard deviations are much more variable than those for the cube method. This makes the rejective algorithm total estimators more variable than those obtained with the cube method.

The conditional variance estimator for fishing effort $\hat{Y}_{ij}$ in site (i, j) proposed in Section 3.2 is

$$v\left(\hat{Y}_{ij}\right) = \frac{H^2\left(1 - n_{ij}/H\right)}{n_{ij}} \sum_{h \in s_{ij}} \frac{\left(y_{hij} - \bar{y}_{sij}\right)^2}{n_{ij} - 1}.$$

The conditional sampling properties, given n_{ij}, of this variance estimator were investigated in the Monte Carlo study with $B = 10{,}000$ balanced samples for the three sample designs. For each site and for each sample size n_{ij} the conditional variance $\operatorname{Var}\left(\hat{Y}_{ij} \mid n_{ij}\right)$ and the conditional expectation of the variance estimator $\mathrm{E}\left\{v\left(\hat{Y}_{ij}\right)\right\}$ were evaluated using the Monte Carlo samples for which the sample size for site (i, j) was n_{ij}. The conditional relative bias of the variance estimator, $\mathrm{E}\left\{v\left(\hat{Y}_{ij}\right)\right\} / \operatorname{Var}\left(\hat{Y}_{ij} \mid n_{ij}\right) - 1$, was then calculated. The conditional relative biases were then aggregated by weighting each sample size n_{ij} using its frequency in the 10,000 Monte Carlo samples; the results are in Table 5.1.

In Table 5.1, the aggregated relative biases are less than 3% in absolute value for the three selection algorithms. This validates the conditional variance estimator proposed in Section 3.2 for a single cell of the cross-classified table. The conditional variances of sums such as $\hat{Y}_{ij} + \hat{Y}_{ij'}$ is more complicated as it involves joint selection probabilities; the estimation of these variances is not considered here. See Breidt and Chauvet (2011) for a discussion of variance estimation with the cube method.

Table 5.1
Aggregated conditional bias, in percentage, of the conditional variance estimator $v\left(\hat{Y}_{ij}\right)$ obtained with the cube method and two rejective algorithms (R 5%, R 50%)

		x has a low variation				x has a medium variation			
Sector	**Site**	x	**CM**	R **5%**	R **50%**	x	**CM**	R **5%**	R **50%**
$i = 1$	$j = 1$	3	1	-3	3	3	-1	1	1
	$j = 2$	2	2	-1	-2	2	3	1	-2
	$j = 3$	3	-1	0	1	3	0	-1	0
$i = 2$	$j = 4$	2	-2	2	0	1	1	-1	-2
	$j = 5$	2	1	-1	-1	2	2	2	3
	$j = 6$	2	0	3	-2	2	0	0	-3
$i = 3$	$j = 7$	3	1	-3	2	2	0	-3	-1
	$j = 8$	3	2	1	1	4	0	0	0
	$j = 9$	4	-1	1	-2	5	-2	-1	1

The conclusion of this Monte Carlo investigation is that the rejective algorithm changes the selection probabilities: sites with small importance are under represented in the rejective samples while the cube method is very good at preserving the selection probabilities. Under both algorithms the calibrated estimator for the total of y in a domain is unbiased. Smaller variances are however obtained with the cube algorithm as it gives domain sample sizes that are less variable than the rejective algorithm.

5 Discussion

In the context of creel surveys, balanced sampling techniques such as the cube method or the rejective algorithm are used to ensure a predetermined sample size in small domains of the survey population. The cube method is very effective at doing so especially in complex survey designs with several stages of sampling. It does not change the selection probabilities and it yields domain sample sizes that are very close their target values. The rejective method, on the other hand, changes the selection probabilities slightly and produce domain sample sizes that are more variable. With a large number of constraints, Fuller's rejective sampling scheme is not really applicable as it requires the evaluation and the inversion of a large covariance matrix in (2.3); alternative acceptation criteria for a sample need to be investigated.

Acknowledgements

We thank the Associate Editor and the referees for their constructive comments on the first version of this manuscript. The assistance and the suggestions of Valérie Bujold, Michel Legault and of Hélène Crépeau who participated to the initial phase of this project is gratefully acknowledged. This project benefitted from the financial assistance of the Canada Research Chair in Statistical Sampling and Data Analysis and form a discovery grant (5244/2012) from the Natural Sciences and Engineering Research Council of Canada.

Appendix

Calculation of the joint selection probabilities when $N = 3$

Consider a population of size 3 and let π_1, π_2, and π_3 be the marginal selection probabilities when drawing a sample of size $n = 2$. The joint selection probabilities π_{ij}, $i \neq j = 1, 2, 3$ satisfy

$$\begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} \pi_{12} \\ \pi_{13} \\ \pi_{23} \end{pmatrix} = \begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \end{pmatrix}.$$

Thus

$$\begin{pmatrix} \pi_{12} \\ \pi_{13} \\ \pi_{23} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}^{-1} \begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1 & 1 & -1 \\ 1 & -1 & 1 \\ -1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \end{pmatrix}.$$

Using these equations, the entries of the covariance matrix (4.1) can be evaluated using the stage 1 and the stage 2 selection probabilities.

References

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. (2009). Stratified balanced sampling. *Survey Methodology*, 35, 1, 115-119. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2009001/article/10888-eng.pdf.

Chauvet, G., Haziza, D. and Lesage, É. (2015). Examining some aspects of balanced sampling in surveys. *Statistica Sinica*.

Daigle, G., Crépeau, H., Bujold, V. and Legault, M. (2015). Enquête de la pêche sportive au bar rayé en Gaspésie en 2015. Technical report.

Deville, J.-C., and Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, 91(4), 893-912.

Fuller, W.A. (2009). Some design properties of a rejective sampling procedure. *Biometrika*, 96(4), 933-944.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, 35, 1491-1523.

Hasler, C., and Tillé, Y. (2014). Fast balanced sampling for highly stratified population. *Computational Statistics & Data Analysis*, 74, 81-94.

Hoenig, J.M., Jones, C.M., Pollock, K.H., Robson, D.S. and Wade, D.L. (1997). Calculation of catch rate and total catch in roving surveys of anglers. *Biometrics*, 306-317.

Legg, J.C., and Yu, C.L. (2010). A comparison of sample set restriction procedures. *Survey Methodology*, 36, 1, 69-79. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2010001/article/11249-eng.pdf.

Minnesota Department of Natural Resources (2011). Creel surveys.

Ousmane Ida, I. (2016). L'échantillonnage équilibré par la méthode du cube et la méthode réjective. Master's thesis, Université Laval.

Pollock, K., Jones, C. and Brown, T. (1994). Angler survey methods and their applications in fisheries management. *American Fisheries Society special publication (USA).*

Robson, D., and Jones, C.M. (1989). The theoretical basis of an access site angler survey design. *Biometrics*, 83-98.

Tillé, Y. (2011). *Sampling Algorithms*. New York: Springer.

United States Environmental Protection Agency (1998). *Guidance for Conducting Fish and Wildlife Consumption Surveys*. EPA, Washington, DC.

Vallée, A.-A., Ferland-Raymond, B., Rivest, L.-P. and Tillé, Y. (2015). Incorporating spatial and operational constraints in the sampling designs for forest inventories. *Environmetrics*, 26(8), 557-570.

Survey Methodology, December 2018
Vol. 44, No. 2, pp. 253-267
Statistics Canada, Catalogue No. 12-001-X

Optimizing a mixed allocation

Antoine Rebecq and Thomas Merly-Alpa[1]

Abstract

This article proposes a criterion for calculating the trade-off in so-called “mixed” allocations, which combine two classic allocations in sampling theory. In INSEE (National Institute of Statistics and Economic Studies) business surveys, it is common to use the arithmetic mean of a proportional allocation and a Neyman allocation (corresponding to a trade-off of 0.5). It is possible to obtain a trade-off value resulting in better properties for the estimators. This value belongs to a region that is obtained by solving an optimization program. Different methods for calculating the trade-off will be presented. An application for business surveys is presented, as well as a comparison with other usual trade-off allocations.

Key Words: Sampling; calculation of allocation; optimization; dispersion of weights; Neyman allocation.

1 Introduction

In this article, we present a framework that replicates part of the surveys carried out in official statistics, specifically business surveys, for which the sampling design is most often a one-stage stratified simple random sampling. A design is created to estimate the total $T(y)$ or the mean $\bar{Y}$ of a continuous key variable of interest y, where y_k designates the value of y for individual k^{th} in the population. Survey data are also used to estimate a collection of other variables, which are sometimes decorrelated or anti-correlated with y.

When a stratified design is used, the choice of an allocation generally serves a specific purpose, based on the classic “one objective, one sample” rule. In order to estimate the total $T(y)$ of the variable of interest with maximum precision, the Neyman allocation (1934) can be used. The specific allocations meet a precise need relative to y. Where survey data are used to estimate quantities from other variables, it is desirable that the design used not deteriorate the quality of the estimators. For example, Cochran (1963) and Chatterjee (1967) propose a specific allocation for a collection of variables of interest. However, this does not solve the case of variables that cannot be included in the creation of the sampling design.

If a variable is decorrelated or anti-correlated to the variables used to calculate a specific allocation, it is known that the variance of the estimate of its total can be very strong (for example, see Ardilly, 2006). Therefore, using a proportional allocation, even when auxiliary information is available, can be advantageous. It enables us to be “agnostic” and to avoid constructing a design that will be harmful to estimate certain variables or certain parameters other than totals or means, or to estimate specific domains. We can also refer to Chiodini, Martelli, Manzi and Verrecchia (2010a, 2010b) for a more extensive discussion on the interest of proportional allocation in the trade-off.

We are interested in a certain type of allocation for stratified samplings with H strata (of respective sizes N_h, for which the sum is equal to the size of the population N) of fixed size n. The choice of an

[1] Antoine Rebecq, MODAL’X - Université Paris Nanterre, 200 avenue de la République, 92000 Nanterre, France. E-mail: antoine.rebecq@parisnanterre.fr; Thomas Merly-Alpa, Insee, 88 Avenue Verdier, 92120 Montrouge, France. E-mail: thomas.merly-alpa@insee.fr.

allocation consists in determining a vector $\mathbf{n} = (n_1, \ldots, n_H)$ verifying constraint $\sum_{h=1}^{H} n_h = n.$ We are specifically studying a "mixed" allocation, which consists of a trade-off between the proportional allocation and another specific allocation, responding to a specific need on one or more variables of interest in a survey:

$$\mathbf{n}_\alpha = \alpha \mathbf{n}_{\text{prop}} + (1-\alpha)\, n_{\text{specific}} \tag{1.1}$$

where $\mathbf{n}_\alpha$, $\mathbf{n}_{\text{prop}}$ and $\mathbf{n}_{\text{specific}}$ are vectors of size H and $\alpha \in [0, 1]$. This trade-off allocation corresponds to the ROAUST method (Chiodini, Manzi, Martelli and Verrecchia, 2017) when the specific allocation chosen is the Neyman allocation (1934). Proportional allocation is defined by:

$$n_h = n \frac{N_h}{N}, \;\; h = 1, \ldots, H.$$

The purpose of this article is to propose a method for determining α. As a result, we would like to calculate a parameter that satisfies a certain optimality criterion that we will detail in Section 2.1. In this article, we will not discuss the composition of the strata, a subject that has been widely explored in the literature, such as in Baillargeon, Rivest and Ferland (2007) and Dalenius and Hodges Jr. (1959). Moreover, we are not trying to account for the phenomena of non-response here.

Note that proportional allocation is one that minimizes the dispersion of weights. Choosing a proportional allocation therefore comes down to the more general logic of choosing a design that minimizes the dispersion of design weights. The design of the INSEE master sample was designed with this objective in mind (Christine and Faivre, 2009), in a design-based logic. In a model-based logic, if we seek to estimate parameters (coefficients of the regression line) and the sampling design is non-informative, then constant design weights minimize variance of the estimate (see Davezies and D'Haultfoeuille, 2009 and Solon, Haider and Wooldridge, 2015).

The trade-off allocation involves reconciling two opposite objectives: creating an effective sampling design for a variable of interest, while keeping the weights as close as possible so as not to deteriorate estimation on very diverse variables. In the following, we will formalize the optimization program corresponding to these constraints. We will present a theorem that will define the criterion of optimality that we seek to resolve. Finally, we will analyze the performances of the allocation determination method that we are proposing and compare them with some other existing methods in the literature on a practical case, particularly a survey of businesses conducted by INSEE (French National Institute of Statistics and Economic Studies).

Several known allocations are already used to perform trade-offs between several objectives. An allocation frequently used at INSEE is a Neyman allocation under local precision constraints, presented in Koubi and Mathern (2009). A better-known allocation in the literature is the Bankier power allocation (1988), which makes a trade-off between the Neyman allocation and an allocation that produces a consistent coefficient of variation of the estimate of the total of a variable of interest on each stratum. This allocation is written as follows:

$$n_h = n \frac{S_h (T_h (x))^q \big/ \bar{Y}_h}{\sum_h S_h (T_h (x))^q \big/ \bar{Y}_h}, \quad h = 1, \ldots, H$$

where q is a parameter in $[0; 1]$, $T_h (x)$ is a measure of the size or importance of stratum h (for example, the size of the stratum or its economic importance), S_h^2 is the empirical variance of y in stratum h and $\bar{Y}_h$ its mean.

In the expression of the Bankier allocation, q is a parameter that, like parameter α of the allocation we are proposing, arbitrates between the two contrary objectives of the allocation: when q is close to 1, the allocation is very close to a Neyman allocation, but when q tends toward 0, the allocation approaches an allocation guaranteeing equal coefficients of variation in all strata. However, the article by Bankier (1988) does not propose a method for choosing this parameter; we will present such a method in this article for our family of mixed allocations.

In this article, we propose to accomplish this trade-off by solving the following program:

$$\min_{\mathbf{n}=(n_h)_{h\in[\![1,\,H]\!]}} \sum_{h=1}^{H} n_h \left(\frac{N_h}{n_h} - \frac{N}{n} \right)^2 + \lambda \left\| \mathbf{n} - \mathbf{n}_{\text{specific}} \right\|_p \tag{1.2}$$

with $\lambda \in [0, +\infty[$, $p \geq 1$ and $\|\cdot\|_p$ denoting the standard p of a vector of size H (in this equation, the term on the right represents a distance between the trade-off allocation and the specific allocation chosen for the survey). We also observe that N/n is the average weight for the sampled units. As in a stratified design, the sampling weight for a unit in stratum h is N_h / n_h; the first term of the optimization program therefore corresponds to the mean square deviation of the weight vector, or the weight dispersion. This program therefore corresponds to a trade-off between the two desired objectives. In part 3, we will see that the interest of the method consists of the choice of an adapted value λ; this choice is decisive for finding the most appropriate balance between the two contrary objectives we are targeting with the allocation, i.e., optimality for certain variables brought by the specific allocation and by equal weighting.

The optimization program used for this paper is inspired by the program used in the CURIOS algorithm (*Curios Uses Representativity Indicators to Optimize Samples*, Merly-Alpa and Rebecq, 2015), which performs an arbitration to establish a prioritization operation for the collection of face-to-face surveys by determining a second-wave allocation. In this paper, we will consider only the problem of determining *ex ante* allocations, and therefore we will not use the algorithm in the context of its introduction.

In Section 2, we present the optimization program that solves the satisfaction of these constraints. In Section 3, we explain how the crucial λ parameter should be chosen. In Section 4, we present a practical application of the mixed allocation on data from French businesses. We conclude in Section 5 by discussing how we could extend the mixed allocation to other designs than the Neyman allocation.

2 Optimization program

The program (1.2) is difficult to resolve and analyze, which is why we will simply look for a solution on a segment between the proportional allocation and a given specific allocation, the Neyman allocation, the one most frequently used. Often, the choice of an $\alpha = 1/2$ is a good trade-off. For example, this is proposed in Chiodini et al. (2010a), or in some INSEE business survey designs.

This method combines the benefits of both methods at a low cost. However, we can question the arbitrary choice of the factor $1/2$. In this paragraph, we will present a method based on a minimization program involving the dispersion of weights as well as the distance to the Neyman allocation to choose a parameter α such as the "optimal" mixed allocation between proportional allocation and the Neyman allocation:

$$\mathbf{n}_{\alpha}^{\text{opt}} = \alpha \mathbf{n}_{\text{prop}} + (1-\alpha)\, \mathbf{n}_{\text{Neyman}}. \tag{2.1}$$

We situate ourselves here in the context of stratified sampling with H strata, ignoring the influence of non-response. This could be integrated by considering anticipated response rates or a second Poisson phase, but this unnecessarily complicates the form of the results. We will focus here on a set of allocations $(\mathbf{n}_{\alpha})$ that go through a segment between the proportional allocation $\left(\mathbf{n}_{\text{prop}}\right)$ and the Neyman allocation $\left(\mathbf{n}_{\text{Neyman}}\right)$, as indicated in equation (2.1). We therefore limit ourselves to achieving the following minimization program, a simplified form of that in equation (1.2):

$$\min_{\alpha \in [0,\,1]} \sum_{h=1}^{H} n_{\alpha,h} \left(\frac{N_h}{n_{\alpha,h}} - \frac{N}{n} \right)^2 + \lambda\alpha. \tag{2.2}$$

The term on the right corresponds to the distance between the desired allocation and the Neyman allocation, up to a constant, integrated in λ: this result is shown in Appendix A.

This minimization program depends on the chosen constant $\lambda \geq 0$. It is clear that when λ is large enough, the term of distance becomes preponderant and we obtain $\alpha = 0$ and therefore $\mathbf{n}_{\alpha} = \mathbf{n}_{\text{Neyman}}$. Similarly, when λ tends toward 0, the factor representing the dispersion of weights becomes preponderant and the allocation tends toward the proportional allocation.

3 Choosing λ

As mentioned in part 1, we must choose an adapted value for λ, which represents the importance we want to give to each term of the trade-off. We will see in this part that the choice of this value is crucial for obtaining a good trade-off parameter. For this, we will focus on the variance of the Horvitz-Thompson estimator of the total of a survey variable of interest obtained with a given allocation when the sampling design applied in each stratum is a simple random sampling.

The idea is to use a key property of the Neyman allocation, which is its flatness (for example, see Ardilly, 2006). This means that in the vicinity of the allocation, the variance of the estimator of the total for the survey variable of interest is close to its minimum value, which is satisfactory from both a theoretical and an empirical standpoint. The issue is properly defining this vicinity: if we succeed in choosing a value λ with which we can produce an allocation sufficiently close to the proportional allocation while belonging to the flat area around the Neyman allocation, we will have succeeded in obtaining weights guaranteeing an estimate with near-optimal precision for the survey's key variable of interest with minimal weight dispersion.

Actually, the choice of λ and the choice of α are interchangeable. For a fixed value of $\lambda,$ we can solve the optimization program of equation (2.2) and obtain a value $\alpha(\lambda),$ and therefore an allocation $\mathbf{n}_{\alpha(\lambda)}.$ Conversely, directly choosing a value of α favours one of the two aspects of the optimization program (distance to the Neyman allocation or equal weighting), similar to the choice of $\lambda.$ Choosing to conserve parameter λ maintains a broader application framework.

We will focus on the variance of the estimator obtained when λ varies. From the allocation $\mathbf{n}_{\alpha(\lambda)},$ it is possible to study the variance of the Horvitz-Thompson estimator of the total of a variable of interest $\hat{T}_{\text{HT}}(y) = \sum_{i \in s} \frac{y_i}{\pi_i},$ with π_i the probability of inclusion in the sample of unit i (i.e., equal to n_h / N_h if h is the stratum that contains i). We then show that there is a "flat" region in the vicinity of the precision optimum (that is, the Neyman allocation, obtained when $\lambda \to +\infty$). Therefore, choosing a λ on this flat region ensures that the precision is only slightly deteriorated from the optimum, while significantly reducing the variance of the survey weights. Mathematically, it is a matter of choosing as λ the torsion point of the curve, whose existence is ensured by the following theorem:

Theorem 1. *Let $V(\lambda)$ be the variance function of $\hat{T}_{\text{HT}}(y)$ for the allocation obtained for the α solution of the minimization program of equation (2.2) for such a $\lambda.$ Therefore, there exists a segment $S \subset [0, +\infty[$ such that:*

- *$\alpha(S) = [0, 1],$ where $\alpha(\lambda)$ associates with λ the solution of program 2.2.*
- *$V(\lambda)$ is decreasing over $S.$*
- *Its second derivative admits a maximum in $S,$ which we call the torsion point.*

This theorem is illustrated in Appendix B.

We therefore want to take λ at the torsion point of the curve, which is also a point of inflection of its derivative; this amounts to situating the "elbow" of the curve, that is, being right at the limit of the variance plateau due to the proximity of the Neyman allocation, linked to the flatness of the optimum. The intuition that justifies this choice is that, on the one hand, the variance of the estimator of the total of the key variable of interest used to calculate the Neyman allocation decreases when λ increases, because the mixed allocation then approaches the Neyman allocation, and on the other hand, beyond a certain threshold, this variance varies little and is very close to its limit, the variance obtained with the Neyman allocation. This threshold corresponds intuitively to the moment when the variance ceases to decrease significantly when λ increases. This point, whose existence is proven by the theorem, is adequately identified by analyzing the variations of the evolutions of the variance with $\lambda,$ i.e., by studying the second derivative of the variance as a function of λ and the point where this derivative reaches a maximum, the derivatives of the variance being negative. Moreover, placing ourselves at the edge of the plateau allows us to limit the maximum of the value of λ and therefore the dispersion of weights. Figure 3.1 illustrates this choice.

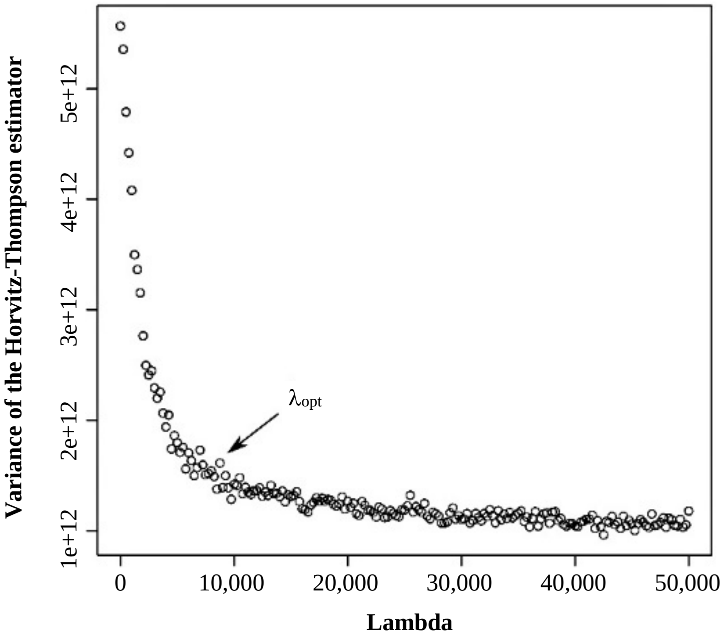


Figure 3.1 Example of a torsion point of the function $V(\lambda)$ for a sampling design explained in Merly-Alpa and Rebecq (2015).

In the simplest cases, meaning once all the information is available, and when sampling takes place in one stage, without considering other parameters, the simplest solution to determine the λ is to analytically study the curve of Figure 3.1 using the classic variance calculation formulas of $\hat{T}_{\text{HT}}(y)$ in a stratified sampling design. The torsion point is obtained by searching for the maximum of the second derivative of the curve $V(\lambda)$. This derivative is generally difficult to calculate analytically, but it is quite possible to find a numerical maximum when we have an analytical formula (or, failing that, a sufficiently smooth curve) for $V(\lambda)$.

Unfortunately, it is not always possible to analytically calculate the variance, such as when other constraints (combining strata, etc.) come into play, or if all the information is not available at the sampling stage. In this case, we replace the curve $V(\lambda)$ with a version estimated by Monte Carlo method:

1. We choose λ in $[0, 1]$.
2. The available data are used to simulate the variance of $\hat{T}_{\text{HT}}(y)$. For this, we calculate the allocation resulting from equation (2.2) for λ and we perform K independent sample draws

based on this sampling design. For each of the $k = 1, \ldots, K$ draws performed, we calculate $\hat{T}_{\mathrm{HT}}^{(k)}(y)$, the Horvitz-Thompson estimator of $T(y)$ obtained with the data from sample k. Next, we calculate the quantity:

$$V_{\mathrm{MC}}(\lambda) = \frac{1}{K-1} \sum_{k=1}^{K} \left(\hat{T}_{\mathrm{HT}}^{(k)}(y) - \frac{1}{K} \sum_{j=1}^{K} \hat{T}_{\mathrm{HT}}^{(j)}(y) \right)^2 .$$

This quantity is a Monte Carlo estimator of $V(\lambda)$.

Note that these simulations require a proxy variable of the variable of interest available for all population units. For business surveys, the turnover available in the tax bases can be a good substitute variable for the actual turnover.

3. We restart for other values of λ covering $[0, 1]$ with a certain step η. The values of η and K should be chosen by considering the calculation time, which can be quite long depending on the original population, but also to ensure that the variance due to the simulations is not too great, which would invalidate the results obtained.
4. Once these results are obtained for different values of λ, we plot the curve of $V_{\mathrm{MC}}(\lambda)$, which we hope is sufficiently smooth. We can then display the curve and visually place the elbow, which allows us to choose the final value of λ_{MC}. Another possibility is to search for the maximum of the second derivative of $V_{\mathbf{MC}}(\lambda)$ using an optimization algorithm sufficiently robust to noise. For example, the algorithm of Nelder and Mead (1965) is implemented in the vast majority of optimization software (e.g., in R or in Python), and Rebecq and Merly-Alpa (2015) show that it gives good practical results for this type of problem.

In all cases, if determining λ at the elbow is difficult, a value should be chosen that ensures that we are to the right of the actual elbow on the curve. This conservative method ensures that we are on the flat region of the curve and that the precision of the estimator of the variable of interest is not impaired.

4 Practical application

We are interested in drawing a sample of 1,000 businesses in the industry based on different stratified sampling designs to learn the total turnover of the sector. The exact field is defined as follows:

- Active businesses located in France.
- Businesses with a workforce between 1 and 100.
- Businesses whose activity sector, measured using the principal activity code, belongs to one of the industry divisions in the Statistical classification of economic activities in the European Community (NACE, whose divisions are identical to the 88 divisions of the International Standard Industrial Classification of All Economic Activities–called ISIC, or CITI in French),

i.e., in divisions 10 to 33, except 12 (Manufacture of tobacco products) and 19 (Manufacture of coke and refined petroleum products), which have a structure too atypical for our study.

The initial population is 102,172 businesses. In general, businesses with a large workforce, i.e., more than 100 employees, are often surveyed exhaustively. Here, we limit ourselves to the non-exhaustive part of a survey.

This population is stratified according to two criteria:

1. The principal activity, at the division level (first two digits).
2. The employee size group, as follows: 1 to 9 employees; 10 to 19 employees; 20 to 49 employees; 50 or more employees.

this constitutes 88 strata, which will be denoted as (A, B), where A is the sector of activity and B the workforce.

We then calculate the proportional and Neyman allocations relative to the dispersion of turnover in each stratum, for $n = 1{,}000.$ Table 4.1 summarizes the characteristics of these two allocations, as well as the strata where the allocation is maximal, both in division 10 (Manufacture of food products).

Table 4.1
Distribution of sample sizes by stratum for both allocations, and sample sizes for strata corresponding to maximum sample sizes

Allocation	Min.	Median	Max.	Stratum	Proportional allocation	Neyman allocation
Proportional	1	3	278	(10, 1-9)	*278*	80
Neyman	1	5	162	(10, 20-49)	18	*162*

We want to choose the optimal mixed allocation for the problem presented in the previous paragraph. For the distance function, we choose the Euclidean distance. Equation 2.2 therefore becomes:

$$\min_{\alpha \in [0,\, 1]} \sum_{h=1}^{H} n_{\alpha,h} \left(\frac{N_h}{n_{\alpha,h}} - \frac{N}{n} \right)^2 + \lambda \sqrt{\sum_{h=1}^{H} \left(n_{\alpha,h} - n_{\text{Neyman},h} \right)^2}. \tag{4.1}$$

We then apply the following method to calculate the optimal allocation:

- Calculate, for different values of $\lambda,$ the value of α solution of the minimization program for equation (4.1).
- For each $\alpha,$ calculate the corresponding allocation.

- For each allocation, analytically calculate the variance of the Horvitz-Thompson estimator for the total turnover. This is possible because we have the turnover of the businesses in the directory used as the survey frame.

The curve represented in Figure 4.1 is finally obtained. We note that its general shape corresponds to what was expected by applying Theorem 1. We visually determine the torsion point, which seems to be located around $1 \cdot 10^7$. So we place $\lambda_{\text{elbow}} = 1 \cdot 10^7$, which is slightly to the right of the elbow, on the flat part of the curve $V(\lambda)$.

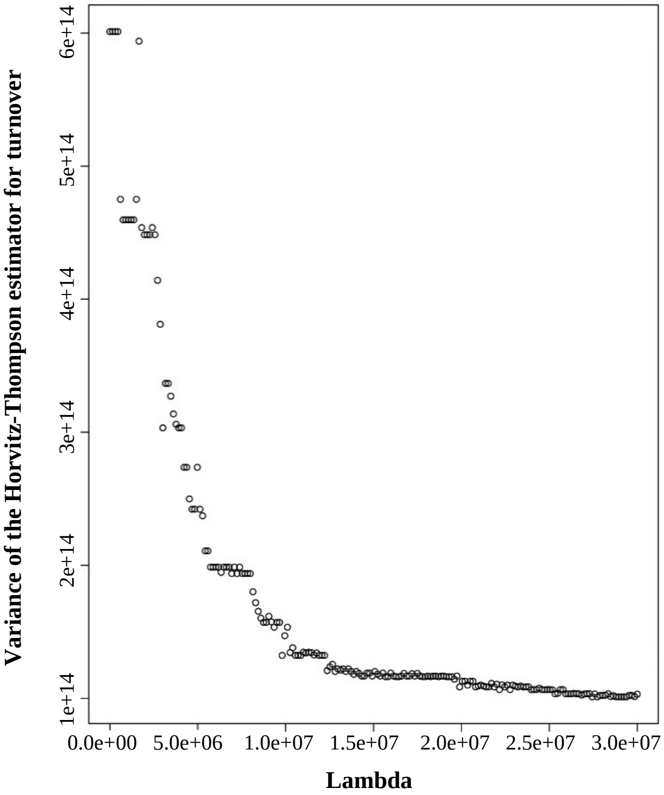


Figure 4.1 Variance of the Horvitz-Thompson estimator for total turnover as part of a trade-off with the Neyman allocation.

We can then use the value of λ_{elbow} to determine α_{elbow}, using the optimization program of equation (4.1). Here, we obtain $\alpha_{\text{elbow}} = 0.644$. This value of α can be interpreted directly. It is close enough to 0.5, which shows that the final allocation is also close enough to what is called the classically mixed allocation, but it is greater than 0.5, which shows that the program optimum is significantly approaching the proportional allocation. The allocation obtained is described in Table 4.2 and is compared with the usual mixed allocation using the arithmetic mean between the two initial allocations.

Table 4.2
Distribution of sample sizes by stratum for the allocation obtained, and for the two strata corresponding to the maximum sample sizes for the Neyman allocation and proportional allocation

Allocation	Min.	Median	Max.	α	Stratum (10, 1-9)	Stratum (10, 20-49)
Proportional	1	3	278	*1*	278	18
Elbow	1	4	208	*0.644*	208	69
Mixed	1	4	179	*0.5*	179	90
Neyman	1	3	162	*0*	80	162

In terms of sample sizes in the strata for the various allocations, we can see that a maximum is obtained for the same stratum as the proportional allocation (10, 1-9), but with a less extensive distribution. Furthermore, stratum (10, 20-49), which has the largest workforce in the Neyman allocation, actually increases in size relative to the proportional allocation, but still remains well below the Neyman allocation. We see the appearance of a trade-off between the allocations, as in the usual mixed allocation.

However, we still have to look at the two criteria that motivate this analysis, namely the standard deviation of the Horvitz-Thompson estimator for the total turnover (in billions of euros), and the dispersion of weights and its influence on the precision of estimators related to other concepts: to evaluate it, we introduce a variable z that is not correlated to turnover. Here, we choose the variable z related to the geographic location of the business defined as follows:

$$z_i = \begin{cases} 1 & \text{if the company } i \text{ is located in Ile-de-France} \\ 0 & \text{otherwise.} \end{cases}$$

We will use these three criteria to compare our method with the initial allocations (proportional, Neyman), but also with the classic mixed allocation (with a factor of 0.5), with Bankier power allocations (1988) for different values of q (where $T_h(\alpha)$ is taken as the sum of turnover in stratum h) and with the Neyman allocation under the local precision constraints from Koubi and Mathern (2009). The results obtained are presented in Table 4.3. In this table, $\hat{T}_{\text{HT}}(\text{CA})$ refers to the Horvitz-Thompson estimator of turnover, and $\hat{T}_{\text{HT}}(z)$ the Horvitz-Thompson estimator of the variable z.

Table 4.3
Dispersion of weights and variance of estimators of turnover and of z for several allocations

Allocation	Parameter	Standard deviation of $\hat{T}_{\text{HT}}(\text{CA})$	Dispersion of weights	Standard deviation of $\hat{T}_{\text{HT}}(z)$
Proportional	$\alpha = 1$	24.7	47	10.7
Elbow	*0.644*	12.5	1,929	11.6
Mixed	*0.5*	11.4	3,473	12.3
Neyman	*0*	9.8	18,585	17.9
Bankier	$q = 0.25$	13.1	36,250	22.2
	0.5	11.2	25,922	19.7
	0.75	10.1	20,187	18.2
Koubi-Mathern	·	12	35,680	22.7

We observe here that the allocation obtained using λ_{elbow} has a precision for the estimate of total turnover that is quite close to the Neyman allocation, while the proportional allocation leads to a much larger standard deviation of the Horvitz-Thompson estimator for total turnover. However, this slight loss in precision is largely offset by the gain in weight dispersion compared with the Neyman allocation and by a significant gain in terms of precision over the total of the geographic variable z. Note that the dispersion of weights is not nil in the proportional allocation because of rounding. When we compare the allocation obtained with the "mixed" strategy using the factor $\alpha = 1/2$, we observe that the loss of a factor 1.1 in the precision of the total turnover is compensated by the gain of a factor 1.8 in weight dispersion and of 1.1 in the precision of the total number of businesses located in the Île-de-France region. The final allocation satisfies our constraints and meets our specification: to have good precision and low dispersion of weights.

Comparison with the methods in the literature illustrates the contribution of the trade-off on the dispersion of weights. For the power allocations, we find that by choosing high values of q corresponding to allocations close to the Neyman allocations, we obtain better precision for the estimate of total turnover than for our allocation. We note that for all the Bankier allocations and for the Neyman allocation under constraints, the weight dispersion is greater than for the Neyman allocation, and therefore much greater than for our allocation. Symmetrically, and as expected, all these allocations contribute to weaken the precision of the estimated total of the variable z.

As the objective of these competing methods is to obtain better local precision, we will examine several subdomains of our field (statistical classification A17 of the French economy):

- Domain C1: Manufacture of food products, beverages;
- Domain C3: Manufacture of electrical, electronic and computer equipment; Manufacture of machines;
- Domain C4: Manufacture of transport equipment;
- Domain C5: Manufacture of other industrial products.

We then compare the precision of the total turnover estimator for each sector. The results are compiled in Table 4.4.

Table 4.4
Local precisions of the total turnover estimator for several allocations

Allocation	**Parameter**	**C1**	**C3**	**C4**	**C5**
Proportional	$\alpha = 1$	0.29	0.30	0.46	0.16
Elbow	*0.644*	0.16	0.20	0.35	0.07
Mixed	*0.5*	0.15	0.18	0.30	0.07
Neyman	*0*	0.12	0.15	0.25	0.06
Bankier	$q = 0.25$	0.21	0.13	0.18	0.07
	0.5	0.17	0.13	0.19	0.06
	0.75	0.14	0.14	0.22	0.06
Koubi-Mathern	·	0.11	0.11	0.11	0.09

We observe here that the allocation we propose gives slightly worse results than the classic mixed allocation on the local precision of the total turnover estimator. However, it is much better than the proportional allocation and, slightly less so, less effective than the Neyman allocation. Our method of choosing α is thus an effective trade-off for reducing the dispersion of weights without overly impacting the overall and local precision of the estimators.

In contrast, and as expected, the allocations with the trade-off objective of maximizing or standardizing local precision are better than the proposed allocation for the majority of the sectors. Therefore, choosing between the trade-off we propose and the one proposed by Bankier (1988) comes down to choosing between better precision for variables not correlated with the variable of interest y (via weight dispersion), like variable z defined here, for our family of mixed allocations, or choosing better local precision for only this variable y in the case of the power allocation. However, the advantage of our method is being able to propose a value of the optimal trade-off parameter α on a certain criterion, which the Bankier method does not do with parameter q.

5 Conclusion

For the stratified designs, we have studied a trade-off allocation situated on a segment between the proportional allocation and the Neyman allocation. A theorem guarantees the existence of a flat region in the vicinity of the optimum and of a particular point that gives an optimal trade-off parameter according to a certain criterion. As part of a survey of businesses in the industry, simulations are conducted showing how the calculation can be done in practice and that the usual choice of a parameter of $1/2$ is not always the most effective. A comparison with other trade-off allocations, such as the classic Bankier allocation, shows that our weight dispersion goal produces more equal weighting at the expense of lower precision for the variable of interest on subdomains of the field. However, it illustrates the variability of the results obtained for the trade-off allocations according to the value of the parameter used; our method for determining parameter α remedies this problem often encountered in the study of these allocation families.

It is possible to replace the Neyman allocation in the trade-off with other specific ad hoc allocations. We postulate that the method remains applicable to obtain the same desirable properties. Different applications of this work were carried out at INSEE with other specific allocations. In the case of the annual Survey on the cost of labour and wage structure (ECMOSS), the specific allocation used for drawing the surveyed businesses is part of a two-stage design where, in each establishment sampled in the first stage, a sample of employees is drawn. The allocation used in the first stage is then optimized to obtain the lowest estimate variance on the estimated total net pay on the final sample of employees, given the dispersion of wages in each establishment. The allocation also integrates precision constraints on certain dissemination domains. Curves of the desired shape are still obtained and the trade-off allocation can be implemented.

Appendix A

Distance term in equation (2.2)

The choice of distance (i.e., a value for p) in the second term of optimization program (1.2) is not crucial in the proposed context, because we will be able to rewrite the second term as follows where C_p is a strictly positive constant dependent only on the choice of p:

$$\left\| \mathbf{n}_{\alpha} - \mathbf{n}_{\text{Neyman}} \right\|_p = \alpha C_p. \tag{A.1}$$

Let us demonstrate this result. By definition (2.1), we have in each stratum h:

$$n_{\alpha,h} = \alpha n_{\text{prop},h} + (1-\alpha)\, n_{\text{Neyman},h}$$

and therefore,

$$n_{\alpha,h} - n_{\text{Neyman},h} = \alpha\left(n_{\text{prop},h} - n_{\text{Neyman},h}\right).$$

We therefore have for any choice of p:

$$\begin{aligned}
\left\| \mathbf{n}_{\alpha} - \mathbf{n}_{\text{Neyman}} \right\|_p &= \left(\sum_{h=1}^{H} \left| n_{\alpha,h} - n_{\text{Neyman},h} \right|^p\right)^{\frac{1}{p}} \\
&= \left(\sum_{h=1}^{H} \alpha^p \left| \left(n_{\text{prop},h} - n_{\text{Neyman},h}\right) \right|^p\right)^{\frac{1}{p}} \\
&= \alpha\left(\sum_{h=1}^{H} \left| n_{\text{prop},h} - n_{\text{Neyman},h} \right|^p\right)^{\frac{1}{p}} \\
&= \alpha C_p.
\end{aligned}$$

We will then integrate C_p, a strictly positive constant, into λ.

Appendix B

Demonstration of Theorem 1

For a $\lambda \geq 0$, the minimization function of program (2.2) is written as follows:

$$\begin{aligned}
f(\alpha) &= \sum_{h=1}^{H} n_{\alpha,h}\left(\frac{N_h}{n_{\alpha,h}} - \frac{N}{n}\right)^2 + \lambda\alpha \\
&= \sum_{h=1}^{H} \left(\frac{N_h^2}{n_{\alpha,h}} - 2\frac{N}{n} N_h + \frac{N^2}{n^2} n_{\alpha,h}\right) + \lambda\alpha \\
&= \sum_{h=1}^{H} \frac{N_h^2}{n_{\alpha,h}} - 2\frac{N^2}{n} + \frac{N^2}{n} + \lambda\alpha \\
&= \sum_{h=1}^{H} \frac{N_h^2}{\alpha\frac{nN_h}{N} + (1-\alpha)\, n_{\text{Neyman},h}} - \frac{N^2}{n} + \lambda\alpha \\
&= \sum_{h=1}^{H} \frac{N_h}{\alpha\frac{n}{N} + (1-\alpha)\frac{n_{\text{Neyman},h}}{N_h}} - \frac{N^2}{n} + \lambda\alpha.
\end{aligned}$$

We now pose for all $h \leq H$:

$$\beta_h = \frac{n}{N} - \frac{n_{\text{Neyman},h}}{N_h}.$$

For each stratum, the β_h represent the difference between the uniform and Neyman sampling fractions. When $\beta_h < 0,$ this means that the Neyman allocation is greater than the proportional allocation; the variable of interest is more dispersed in this stratum. Let us now derive f:

$$f'(\alpha) = \sum_{h=1}^{H} \frac{-N_h \beta_h}{\left(\alpha\beta_h + \frac{n_{\text{Neyman},h}}{N_h}\right)^2} + \lambda. \tag{B.1}$$

We deduce from equation (B.1) that the derivative cancels out when:

$$\lambda = \sum_{h=1}^{H} \frac{N_h \beta_h}{\left(\alpha\beta_h + \frac{n_{\text{Neyman},h}}{N_h}\right)^2} =: g(\alpha).$$

Now function g_h defined as follows:

$$g_h : \alpha \to \frac{N_h \beta_h}{\left(\alpha\beta_h + \frac{n_{\text{Neyman},h}}{N_h}\right)^2}$$

is decreasing. So:

- If β_h is negative, the denominator decreases when α increases. In this case, its inverse increases with α. Therefore, we multiply by β_h to obtain g_h, which implies that g_h is decreasing.
- If β_h is positive, the denominator increases when α increases. By inverting and then multiplying by β_h, we find that g_h is decreasing.

So, if $\lambda \in [g(1), g(0)]$, we know that there is an α_0 that cancels the derivative. As f' evolves inversely to g, f' is increasing and therefore α_0 is the minimum of f on $[0, 1]$.

Furthermore, as $g(\alpha_0) = \lambda$ by definition, the decrease of g implies that when λ increases in $[g(1), g(0)]$, then α_0 decreases. We therefore use the following lemma, admitted because it is relative to a classic property of the Neyman allocation:

Lemma 1. *The function that at α associates the variance of the Horvitz-Thompson estimator of the variable of interest X for the allocation $n_{\alpha,h}$ is increasing.*

We deduce that $V(\lambda)$ is decreasing over S. Finally, by continuity, V'' admits a maximum over S.

References

Ardilly, P. (2006). *Les Techniques de Sondage*. Editions Technip.

Baillargeon, S., Rivest, L.-P. and Ferland, M. (2007). Stratification en enquêtes entreprises : une revue et quelques avancées. *Proceedings of the Survey Methods Section, Statistical Society of Canada*.

Bankier, M.D. (1988). Power allocations: Determining sample sizes for subnational areas. *The American Statistician*, 42.3, 174-177.

Chatterjee, S. (1967). A note on optimum allocation. *Scandinavian Actuarial Journal*, 1-2, 40-44.

Chiodini, P.M., Manzi, G., Martelli, B.M. and Verrecchia, F. (2017). *Sampling Allocation Strategies: A Simulation-Based Comparison*. url: https://events.unibo.it/itacosm2017/abtracts-of-invited-papers/manzi-et-al_presentation_itacosm17.pdf/@@download/file/Manzi%20et%20al_Presentation_ ITACOSM17.pdf.

Chiodini, P.M., Martelli, B.M., Manzi, G. and Verrecchia, F. (2010a). The ISAE manufacturing survey sample: Validating the Nace Rev. 2 sectorial allocation. *Economic Tendency Surveys and the Services Sector*. CIRET.

Chiodini, P.M., Martelli, B.M., Manzi, G. and Verrecchia, F. (2010b). Between theoretical and applied approach: Which compromise for unit allocation in business surveys? *SIS Conference*. Società italiana di statistica.

Christine, M., and Faivre, S. (2009). Le projet OCTOPUSSE de nouvel Échantillon-Maître de l’INSEE. *JMS*, 24.

Cochran, W. (1963). *Sampling Techniques*.

Dalenius, T., and Hodges Jr, J.L. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, 54.285, 88-101.

Davezies, L., and D’Haultfoeuille, X. (2009). Faut-il pondérer ?... Ou l’éternelle question de l’économètre confronté à un problème de sondage. Working paper of INSEE.

Koubi, M., and Mathern, S. (2009). Résolution d’une des limites de l’allocation de Neyman. *JMS*, 1.

Merly-Alpa, T., and Rebecq, A. (2015). L’algorithme CURIOS pour l’optimisation du plan de sondage en fonction de la non-réponse. *Journées de la Statistique de la SFdS*, Lille.

Nelder, J.A., and Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, 7.4, 308-313.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 558-625.

Rebecq, A., and Merly-Alpa, T. (2015). Algorithme CURIOS et méthode de « priorisation » pour les enquêtes en face-à-face. Application à l’enquête Patrimoine 2014. *Actes des Journées de Méthodologie Statistique*.

Solon, G., Haider, S.J. and Wooldridge, J.M. (2015). What are we weighting for? *Journal of Human Resources*, 50.2, 301-316.

Survey Methodology, December 2018
Vol. 44, No. 2, pp. 269-289
Statistics Canada, Catalogue No. 12-001-X

Variance estimation under monotone non-response for a panel survey

Hélène Juillard and Guillaume Chauvet[1]

Abstract

Panel surveys are frequently used to measure the evolution of parameters over time. Panel samples may suffer from different types of unit non-response, which is currently handled by estimating the response probabilites and by reweighting respondents. In this work, we consider estimation and variance estimation under unit non-response for panel surveys. Extending the work by Kim and Kim (2007) for several times, we consider a propensity score adjusted estimator accounting for initial non-response and attrition, and propose a suitable variance estimator. It is then extended to cover most estimators encountered in surveys, including calibrated estimators, complex parameters and longitudinal estimators. The properties of the proposed variance estimator and of a simplified variance estimator are estimated through a simulation study. An illustration of the proposed methods on data from the ELFE survey is also presented.

Key Words: Longitudinal estimation; Non-response model; Product sampling design; Response homogeneity groups; Simplified variance estimation.

1 Introduction

Surveys are not only used to produce estimators for one point in time (cross-sectional estimations), but also to measure the evolution of parameters (longitudinal estimations), and are thus repeated over time. In this paper, we are interested in estimation and variance estimation for panel surveys, in which measures are repeated over time for units in a same sample (Kalton, 2009). Among the panel surveys (also known as longitudinal surveys, see Lynn, 2009), cohort surveys are particular cases where the units in the sample are linked by a common original event, such as being born on the same year for children in the ELFE survey (Enquête longitudinale française depuis l'enfance), which is the motivating example for this work.

ELFE is the first longitudinal study of its kind in France, tracking children from birth to adulthood (Pirus, Bois, Dufourg, Lanoë, Vandentorren, Leridon and the Elfe team, 2010). Covering the whole metropolitan France, it was launched in 2011 and consists of more than 18,000 children whose parents consented to their inclusion. It will examine every aspect of these children's lives from the perspectives of health, social sciences and environmental health. The ELFE survey suffers from unit non-response, which needs to be accounted for by using available auxiliary information, so as to limit the bias of estimators. Though the ELFE survey will be used for illustration in this paper, non-response occurs in virtually any panel survey so that the proposed methods are of general interest; see for example Laurie, Smith and Scott (1999) for the treatment of non-response of the British Household Panel Survey, or Vandecasteele and Debels (2007) for the European Community Household Panel.

Non-response is currently handled by modeling the response probabilities (Kim and Kim, 2007) and by reweighting respondents with the inverse of these estimated probabilities, which leads to the so-called

[1] Hélène Juillard, INED, 133 boul. Davout, 75020 Paris, France; Guillaume Chauvet, ENSAI/IRMAR, Campus de Ker Lann, 35170 Bruz, France. E-mail: guillaume.chauvet@ensai.fr.

propensity score adjusted estimator. A panel sample may suffer from three types of unit non-response (Hawkes and Plewis, 2009): initial non-response refers to the original absence of selected units; wave non-response occurs when some units in the panel sample temporarily do not answer at some point in time, while attrition occurs when some units in the panel sample permanently do not answer from some point in time. Wave non-response was fairly uncommon in the first waves of the ELFE survey which were at our disposal. We therefore simplify this set-up by assuming monotone non-response, where only initial non-response and attrition occur.

There is a vast literature on the treatment of unit non-response for surveys over time, see Ekholm and Laaksonen (1991), Fuller, Loughin and Baker (1994), Rizzo, Kalton and Brick (1996), Clarke and Tate (2002), Laaksonen and Chambers (2006), Hawkes and Plewis (2009), Rendtel and Harms (2009), Laaksonen (2007), Slud and Bailey (2010), Zhou and Kim (2012). Variance estimation for longitudinal estimators is considered in Tam (1984), Laniel (1988), Nordberg (2000), Berger (2004), Skinner and Vieira (2005), Qualité and Tillé (2008) and Chauvet and Goga (2018), but with focus on the sampling variance only. Variance estimation in case of non-response weighting adjustments on cross-sectional surveys is considered in Kim and Kim (2007). To the best of our knowledge, and despite the interest for applications, variance estimation accounting for non-response for panel surveys has not been treated in the literature, with the exception of Zhou and Kim (2012).

Zhou and Kim (2012) consider the estimation of a mean for a panel survey, in case of monotone non-response. Instead of using the propensity score adjusted estimator, Zhou and Kim (2012) define an optimal propensity score estimator. It is obtained by noting that for any variable of interest observed before time t, the estimator produced at time t differs from the estimator obtained at the date when the variable was observed, which is based on a larger sample. Adjusting on these differences by means of some form of calibration leads to the estimator proposed by Zhou and Kim (2012). It makes full use of the information collected at previous times, and it is therefore expected to be more efficient than the propensity score adjusted estimator. However, a panel survey may include a large number of variables of interest observed at several times, and calibrating on a too large number of variables may lead to estimators whose performances are worsened (Silva and Skinner, 1997). A careful modeling exercise seems therefore necessary before applying the optimal estimator of Zhou and Kim (2012). In this work, we rather focus on the propensity score adjusted estimator, which is popular in practice.

Zhou and Kim (2012) also consider variance estimation for their optimal estimator, under the so-called reverse framework of Fay (1992). By viewing the sample obtained at time t as the result of a two-phase process, the first phase being associated to the original sampling design and the second phase to the successive non-response steps, it is assumed under the reverse framework that these two phases may be reversed. This requires the two-phase process to be strongly invariant as defined by Beaumont and Haziza (2016). In this paper, we propose a general variance estimator for the propensity score adjusted estimator, for which the strong invariance assumption is not needed. We also extend this variance estimator to account for estimation of complex parameters, possibly with calibrated weights, and to cover longitudinal estimators.

In each case, a simplified conservative variance estimator, which may be easier to compute for secondary users, is also proposed.

The paper is organized as follows. In Section 2, we first define the notation. A parametric model is then postulated, leading to estimated response probabilities and to a reweighted estimator. A variance estimator is then derived by following the approach in Kim and Kim (2007), and a simplified version is also proposed. They are illustrated in the particular case of the logistic regression model. The proposed variance estimator is extended to cover calibrated estimators and complex parameters in Section 3. Longitudinal estimation is discussed in Section 4, and the proposed variance estimator is used to cover such cases. The variance estimators are compared in Section 5 through a simulation study, and an illustration on the ELFE data is proposed in Section 6. We draw some conclusions in Section 7.

2 Correction of non-response and attrition

2.1 Notation and main assumptions

We are interested in a finite population U. A sample s_0 is first selected according to some sampling design $p(\cdot)$, and we assume that the first-order inclusion probabilities π_i are strictly positive for any $i \in U$. This first sampling phase corresponds to the original inclusion of units in the sample.

We consider the case of a panel survey in which the sole units in the original sample s_0 are followed over time, without reentry or late entry units at subsequent times to represent possible newborns. We are therefore interested in estimating some parameter defined over the population U, for some study variable y_t taking the value y_{it} for the unit i at time t. The units in the sample s_0 are followed at subsequent times $\delta = 1, \ldots, t$, and the sample is prone to unit non-response at each time. We note r_i^δ for the response indicator for unit i at time δ, and s_δ for the subset of respondents at time δ.

We assume monotone non-response resulting in the nested sequence $s_0 \supset s_1 \supset \ldots \supset s_t$. For $\delta = 1, \ldots, t$, we note $p_i^\delta = \Pr(i \in s_\delta \,|\, s_{\delta-1})$ for the response probability of some unit i to be a respondent at time δ. We assume that the data are missing at random, i.e. the response probability p_i^δ at time δ can be explained by the variables observed at times $0, \ldots, \delta - 1$, including the variables of interest, see for example Zhou and Kim (2012). Also, we assume that at any time δ the units answer independently of one another, and we note $p_{ij}^\delta = p_i^\delta p_j^\delta$ for the probability that two distinct units i and j answer jointly at time δ.

2.2 Reweighted estimator

We are interested in estimating the total $Y(t) = \sum_{i \in U} y_{it}$ at time t. In practice, the response probabilities at each time are unknown and need to be estimated. We assume that at each time δ the probability of response is parametrically modeled as

$$p_i^\delta = f^\delta\left(z_i^\delta, \alpha^\delta\right) \tag{2.1}$$

for some known function $f^\delta(\cdot,\cdot)$, where z_i^δ is a vector of variables observed for all the units in $s_{\delta-1}$, and α^δ denotes some unknown parameter. Here and elsewhere, the superscript δ will be used when we account for non-response at time δ, like for the probability p_i^δ of unit i to be a respondent at time δ. Following the approach in Kim and Kim (2007), we assume that the true parameter is estimated by $\hat{\alpha}^\delta$, the solution of the estimating equation

$$\frac{\partial}{\partial \alpha} \sum_{i \in s_{\delta-1}} k_i^\delta \left\{ r_i^\delta \ln\left(p_i^\delta\right) + \left(1 - r_i^\delta\right) \ln\left(1 - p_i^\delta\right) \right\} = 0, \tag{2.2}$$

with k_i^δ some weight of unit i in the estimating equation. Customary choices for these weights include $k_i^\delta = 1$ and $k_i^\delta = \pi_i^{-1}$, see Fuller and An (1998), Beaumont (2005) and Kim and Kim (2007).

The estimated response probability at time δ is $\hat{p}_i^\delta = f^\delta\left(z_i^\delta, \hat{\alpha}^\delta\right)$. The propensity score adjusted estimator at time t, which will be simply called the reweighted estimator in what follows, is defined as

$$\hat{Y}_t(t) = \sum_{i \in s_t} \frac{y_{it}}{\pi_i \hat{p}_i^{1\to t}} \qquad \text{with} \qquad \hat{p}_i^{1\to t} = \prod_{\delta=1}^{t} \hat{p}_i^\delta. \tag{2.3}$$

Here and elsewhere, the subscript t will be used when the sample observed at time t is used for estimation, like for $\hat{Y}_t(\cdot)$ which makes use of the sample s_t. We simplify the notation as $\hat{Y}_t(t) \equiv \hat{Y}_t$ when the total at time t is estimated by using the sample observed at time t.

2.3 Variance computation

Under some regularity assumptions on the response mechanisms and some regularity conditions on the $p^\delta(\cdot,\cdot)$'s, we obtain from Theorem 1 in Kim and Kim (2007) that we can write

$$\hat{Y}_t = \hat{Y}_{\text{lin},\,t}(t) + O_p\left(Nn^{-1}\right), \tag{2.4}$$

where

$$\hat{Y}_{\text{lin},\,t}(t) = \sum_{i \in s_{t-1}} \frac{1}{\pi_i \hat{p}_i^{1\to t-1}} \left\{ k_i^t \pi_i \hat{p}_i^{1\to t-1} p_i^t \left(h_i^t\right)^\top \gamma^t + \frac{r_i^t}{p_i^t} \left(y_{it} - k_i^t \pi_i \hat{p}_i^{1\to t-1} p_i^t \left(h_i^t\right)^\top \gamma^t \right) \right\}, \tag{2.5}$$

and where for any $\delta = 1, \ldots, t$ we denote by h_i^δ the value of $h_i^\delta(\alpha) = \partial \text{logit}\left(p_i^\delta\right) / \partial \alpha$ evaluated at $\alpha = \alpha^\delta$, and

$$\gamma^\delta = \left\{ \sum_{i \in s_{\delta-1}} k_i^\delta p_i^\delta \left(1 - p_i^\delta\right) h_i^\delta \left(h_i^\delta\right)^\top \right\}^{-1} \sum_{i \in s_{\delta-1}} \frac{1 - p_i^\delta}{\hat{p}_i^{1\to\delta-1}} h_i^\delta \frac{y_{it}}{\pi_i}. \tag{2.6}$$

From (2.5), we obtain that

$$E\left\{ \hat{Y}_{\text{lin},\,t}(t) \middle|\, s_{t-1} \right\} = \hat{Y}_{t-1}(t), \tag{2.7}$$

with $\hat{Y}_{t-1}(t)$ the estimator of $Y(t)$ computed on s_{t-1}. Using a proof by induction, it follows from (2.4) and (2.7) that $\hat{Y}_t$ is approximately unbiased for $Y(t)$. Also, the variance of $\hat{Y}_t$ may be asymptotically approximated by

$$V_{\text{app}}\left(\hat{Y}_t\right) = V\left(\sum_{i \in s_0} \frac{y_{it}}{\pi_i}\right) + E\left[\sum_{\delta=1}^{t} V\left\{\hat{Y}_{\text{lin},\delta}(t) \mid s_{\delta-1}\right\}\right]. \tag{2.8}$$

The first term in the right-hand side of (2.8) is the variance due to the sampling design, that we note as $V^{p}\left(\hat{Y}_t\right)$. The second term in the right-hand side of (2.8) is the variance due to non-response, that we note as $V^{\text{nr}}\left(\hat{Y}_t\right)$. From (2.5), this asymptotic variance is given by

$$V^{\text{nr}}\left(\hat{Y}_t\right) = E\left(\sum_{\delta=1}^{t} V^{\text{nr}\delta}\left(\hat{Y}_t\right)\right), \tag{2.9}$$

where

$$V^{\text{nr}\delta}\left(\hat{Y}_t\right) = \sum_{i \in s_{\delta-1}} p_i^{\delta}\left(1 - p_i^{\delta}\right)\left(\frac{y_{it}}{\pi_i \hat{p}_i^{1\to\delta-1} p_i^{\delta}} - k_i^{\delta}\left(h_i^{\delta}\right)^{\top} \gamma^{\delta}\right)^2. \tag{2.10}$$

We note that for each of its component $\delta = 1, \ldots, t$, the term $V^{\text{nr}\delta}\left(\hat{Y}_t\right)$ in (2.10) includes a centering term $k_i^{\delta}\left(h_i^{\delta}\right)^{\top} \gamma^{\delta}$, which is essentially a prediction of $\left(\pi_i \hat{p}_i^{1\to\delta-1} p_i^{\delta}\right)^{-1} y_i$ by means of regressors h_i^{δ}. This centering is due to the estimation of the response probabilities. Suppressing these centering terms, equations (2.9) and (2.10) would lead to the variance of the estimator of $Y(t)$ we would obtain by replacing in (2.3) the estimated probabilities by their true values. The variance of this estimator is usually larger than that of the reweighted estimator in (2.3); see also Beaumont (2005), equation (5.7) and Kim and Kim (2007), equation (17), for the case $t = 1$.

2.4 Variance estimation

At time t, an approximately unbiased estimator for the variance due to the sampling design $V^{p}\left(\hat{Y}_t\right)$ is

$$\hat{V}_t^{p}\left(\hat{Y}_t\right) = \sum_{i,j \in s_t} \frac{\Delta_{ij}}{\pi_{ij}} \frac{1}{\hat{p}_{ij}^{1\to t}} \frac{y_{it}}{\pi_i} \frac{y_{jt}}{\pi_j}, \tag{2.11}$$

where $\hat{p}_{ij}^{1\to t} \equiv \prod_{\delta=1}^{t} \hat{p}_{ij}^{\delta}$, and where $\hat{p}_{ij}^{\delta} = \hat{p}_i^{\delta}$ if $i = j$, and $\hat{p}_{ij}^{\delta} = \hat{p}_i^{\delta} \hat{p}_j^{\delta}$ otherwise. Following equation (25) in Kim and Kim (2007), $V^{\text{nr}}\left(\hat{Y}_t\right)$ may be approximately unbiasedly estimated at time t by

$$\hat{V}_t^{\text{nr}}\left(\hat{Y}_t\right) = \sum_{\delta=1}^{t} \hat{V}_t^{\text{nr}\delta}\left(\hat{Y}_t\right) \tag{2.12}$$

where

$$\hat{V}_t^{\text{nr}\delta}\left(\hat{Y}_t\right) = \sum_{i \in s_t} \frac{\hat{p}_i^{\delta}\left(1 - \hat{p}_i^{\delta}\right)}{\hat{p}_i^{\delta\to t}}\left(\frac{y_{it}}{\pi_i \hat{p}_i^{1\to\delta}} - k_i^{\delta}\left(\hat{h}_i^{\delta}\right)^{\top} \hat{\gamma}_t^{\delta}\right)^2, \tag{2.13}$$

$$\hat{h}_i^{\delta} = h\left(z_i, \hat{\alpha}^{\delta}\right), \tag{2.14}$$

$$\hat{\gamma}_t^{\delta} = \left\{ \sum_{i \in s_t} k_i^{\delta} \frac{\hat{p}_i^{\delta} \left(1 - \hat{p}_i^{\delta}\right)}{\hat{p}_i^{\delta \to t}} \hat{h}_i^{\delta} \left(\hat{h}_i^{\delta}\right)^{\top} \right\}^{-1} \sum_{i \in s_t} \frac{1 - \hat{p}_i^{\delta}}{\hat{p}_i^{1 \to t}} \hat{h}_i^{\delta} \frac{y_{it}}{\pi_i}. \tag{2.15}$$

This leads to the global variance estimator at time t

$$\hat{V}_t \left(\hat{Y}_t\right) = \hat{V}_t^{p} \left(\hat{Y}_t\right) + \hat{V}_t^{\mathrm{nr}} \left(\hat{Y}_t\right). \tag{2.16}$$

A simplified estimator of the variance due to non-response is obtained by ignoring the prediction terms $k_i^{\delta} \left(\hat{h}_i^{\delta}\right)^{\top} \hat{\gamma}_t^{\delta}$ for each of the $\delta = 1, \ldots, t$ variance components. After some algebra, this leads to the simplified variance estimator

$$\hat{V}_{t,\,\mathrm{simp}}^{\mathrm{nr}} \left\{\hat{Y}_t\,(t)\right\} = \sum_{i \in s_t} \frac{1 - \hat{p}_i^{1 \to t}}{\left(\hat{p}_i^{1 \to t}\right)^2} \left(\frac{y_{it}}{\pi_i}\right)^2. \tag{2.17}$$

The main advantage of this simplified variance estimator is that it only requires the knowledge of the estimated response probabilities. On the other hand, the computation of the variance estimator in (2.12) requires the knowledge of the response models used at all times. The simplified variance estimator is therefore of particular interest for secondary users of the survey data, for which the estimated response probabilities may be the only available information related to the response modeling. This simplified variance estimator will tend to overestimate the variance due to non-response of $\left(\hat{Y}_t\right)$ if the prediction term $k_i^{\delta} \left(h_i^{\delta}\right)^{\top} \gamma^{\delta}$ partly explains $\left(\pi_i \hat{p}_i^{1 \to \delta - 1} p_i^{\delta}\right)^{-1} y_{it}$.

2.5 Application to the logistic regression model

In the particular case when a logistic regression model is used at each time δ, the model (2.1) may be rewritten as

$$\mathrm{logit}\left(p_i^{\delta}\right) = \left(z_i^{\delta}\right)^{\top} \alpha^{\delta}. \tag{2.18}$$

We obtain $\hat{h}_i^{\delta} = z_i^{\delta}$, and the estimator for the variance due to non-response is given by (2.12), with

$$\hat{V}_t^{\mathrm{nr}\delta} \left(\hat{Y}_t\right) = \sum_{i \in s_t} \frac{\hat{p}_i^{\delta} \left(1 - \hat{p}_i^{\delta}\right)}{\hat{p}_i^{\delta \to t}} \left(\frac{y_{it}}{\pi_i \hat{p}_i^{1 \to \delta}} - k_i^{\delta} \left(z_i^{\delta}\right)^{\top} \hat{\gamma}_t^{\delta} \right)^2, \tag{2.19}$$

$$\hat{\gamma}_t^{\delta} = \left\{ \sum_{i \in s_t} k_i^{\delta} \frac{\hat{p}_i^{\delta} \left(1 - \hat{p}_i^{\delta}\right)}{\hat{p}_i^{\delta \to t}} z_i^{\delta} \left(z_i^{\delta}\right)^{\top} \right\}^{-1} \sum_{i \in s_t} \frac{1 - \hat{p}_i^{\delta}}{\hat{p}_i^{1 \to t}} z_i^{\delta} \frac{y_{it}}{\pi_i}. \tag{2.20}$$

If the reweighted estimator is computed at time $t = 1$, the estimator in (2.12) for the variance due to non-response may be rewritten as

$$\hat{V}_1^{\mathrm{nr}} \left(\hat{Y}_1\right) = \sum_{i \in s_1} \left(1 - \hat{p}_i^{1}\right) \left(\frac{y_{i1}}{\pi_i \hat{p}_i^{1}} - k_i^{1} \left(z_i^{1}\right)^{\top} \hat{\gamma}_1^{1} \right)^2. \tag{2.21}$$

If the reweighted estimator is computed at time $t = 2,$ the estimator in (2.12) for the variance due to non-response may be rewritten as

$$\hat{V}_2^{\mathrm{nr}}\left(\hat{Y}_2\right) = \sum_{i \in s_2} \frac{\left(1 - \hat{p}_i^1\right)}{\hat{p}_i^2}\left(\frac{y_{i2}}{\pi_i \hat{p}_i^1} - k_i^1 \left(z_i^1\right)^\top \hat{\gamma}_2^1\right)^2$$

$$+ \sum_{i \in s_2}\left(1 - \hat{p}_i^2\right)\left(\frac{y_{i2}}{\pi_i \hat{p}_i^1 \hat{p}_i^2} - k_i^2 \left(z_i^2\right)^\top \hat{\gamma}_2^2\right)^2. \qquad (2.22)$$

In practice, the model of Response Homogeneity Groups (RHG) is often assumed when correcting for unit non-response. Under this model, it is assumed that at each time $\delta = 1, \ldots, t,$ the sub-sample $s_{\delta-1}$ may be partitioned into $C(\delta - 1)$ groups $s_{\delta-1}^c, c = 1, \ldots, C(\delta - 1),$ such that the response probability p_i^δ is constant inside a group. This model is a particular case of the logistic regression model in (2.18), obtained with

$$z_i^\delta = \left[1\left\{i \in s_{\delta-1}^1\right\}, \ldots, 1\left\{i \in s_{\delta-1}^{C(\delta-1)}\right\}\right]^\top, \qquad (2.23)$$

and the variance due to non-response is estimated accordingly. Explicit formulas are given in Appendix.

3 Calibration and complex parameters

In most surveys, a calibration step is used to obtain adjusted weights which enable to improve the accuracy of total estimates. Such calibrated estimators are considered in Section 3.1. Also, more complex parameters than totals are frequently of interest, and a linearization step can be used for variance estimation. This is the purpose of Section 3.2. The estimation of complex parameters with calibrated weights is treated in Section 3.3. In each case, explicit formulas for variance estimation and simplified variance estimation are derived, and the bias of the simplified variance estimator is discussed.

3.1 Variance estimation for calibrated total estimators

Assume that a vector x_i of auxiliary variables is available for any unit $i \in s_t,$ and that the vector of totals X on the population U is known. Then an additional calibration step (Deville and Särndal, 1992) is usually applied to $\hat{Y}_t.$ It consists in modifying the weights $d_{ti} = \pi_i^{-1}\left(\hat{p}_i^{1 \to t}\right)^{-1}$ to obtain calibrated weights w_{ti} which enable to match the real total $X,$ in the sense that

$$\sum_{i \in s_t} w_{ti} x_i = X. \qquad (3.1)$$

The new calibrated weights are chosen to minimize a distance function with the original weights, while satisfying (3.1). This leads to the calibrated estimator

$$\hat{Y}_{wt} = \sum_{i \in s_t} w_{ti} y_{it}. \qquad (3.2)$$

The estimated residual for the weighted regression of y_{it} on x_i is denoted by

$$e_{it} = y_{it} - \hat{b}_t x_i \tag{3.3}$$

with

$$\hat{b}_t = \left(\sum_{i \in s_t} \frac{1}{\pi_i \hat{p}_i^{1 \to t}} x_i x_i^\top\right)^{-1} \sum_{i \in s_t} \frac{1}{\pi_i \hat{p}_i^{1 \to t}} x_i y_{it}. \tag{3.4}$$

Replacing in (2.11) the variable y_{it} with e_{it} yields the estimator of the variance due to the sampling design

$$\hat{V}_t^p\left(\hat{Y}_{wt}\right) = \sum_{i,\, j \in s_t} \frac{\Delta_{ij}}{\pi_{ij}} \frac{1}{\hat{p}_{ij}^{1 \to t}} \frac{e_{it}}{\pi_i} \frac{e_{jt}}{\pi_j}. \tag{3.5}$$

Similarly, replacing in (2.12) the variable y_{it} with e_{it} yields the estimator of the variance due to the non-response

$$\hat{V}_t^{\text{nr}}\left(\hat{Y}_{wt}\right) = \sum_{\delta=1}^{t} \sum_{i \in s_t} \frac{\hat{p}_i^\delta\left(1 - \hat{p}_i^\delta\right)}{\hat{p}_i^{\delta \to t}} \left(\frac{e_{it}}{\pi_i \hat{p}_i^{1 \to \delta}} - k_i^\delta \left(\hat{h}_i^\delta\right)^\top \hat{\gamma}_{te}^\delta\right)^2 \tag{3.6}$$

$$\hat{\gamma}_{te}^\delta = \left\{\sum_{i \in s_t} k_i^\delta \frac{\hat{p}_i^\delta\left(1 - \hat{p}_i^\delta\right)}{\hat{p}_i^{\delta \to t}} \hat{h}_i^\delta \left(\hat{h}_i^\delta\right)^\top\right\}^{-1} \sum_{i \in s_t} \frac{1 - \hat{p}_i^\delta}{\hat{p}_i^{1 \to t}} \hat{h}_i^\delta \frac{e_{it}}{\pi_i}. \tag{3.7}$$

The global variance estimator for $\hat{Y}_{wt}$ is

$$\hat{V}_t\left(\hat{Y}_{wt}\right) = \hat{V}_t^p\left(\hat{Y}_{wt}\right) + \hat{V}_t^{\text{nr}}\left(\hat{Y}_{wt}\right). \tag{3.8}$$

The simplified estimator of the variance due to non-response is

$$\hat{V}_{t,\,\text{simp}}^{\text{nr}}\left(\hat{Y}_{wt}\right) = \sum_{i \in s_t} \frac{1 - \hat{p}_i^{1 \to t}}{\left(\hat{p}_i^{1 \to t}\right)^2} \left(\frac{e_{it}}{\pi_i}\right)^2. \tag{3.9}$$

Here again, this simplified variance estimator ignores the prediction terms $k_i^\delta \left(\hat{h}_i^\delta\right)^\top \hat{\gamma}_{te}^\delta$. If the underlying calibration model is appropriate, then the explanatory power of $\hat{h}_i^\delta$ for e_{it} is expected to be small, as well as the bias of the simplified variance estimator. On the other hand, if there remains in e_{it} some significant part of y_{it} that may not been explained by x_i, the bias of the simplified variance estimator may be non-negligible. This may occur in case of domain estimation, when the calibration variables do not include any auxiliary information specific of the domain.

3.2 Variance estimation for complex parameters

We may be interested in estimating more complex parameters than totals. Suppose that the variable of interest y_{it} is $q-$multivariate, and that the parameter of interest is $\theta(t) = f\{Y(t)\}$ with $f(\cdot)$ a known function. At time t, substituting $\hat{Y}_t$ into $\theta(t)$ yields the plug-in estimator $\hat{\theta}_t = f\left(\hat{Y}_t\right)$.

The estimated linearized variable of $\theta(t)$ is

$$u_{it} = \left\{f'\left(\hat{Y}_t\right)\right\}^\top y_{it}, \tag{3.10}$$

with $f'(\hat{Y}_t)$ the $q-$vector of first derivatives of f at point $\hat{Y}_t$. Replacing in (2.11) the variable y_{it} with u_{it} yields the estimator of the variance due to the sampling design

$$\hat{V}_t^p\left(\hat{\theta}_t\right) = \sum_{i,\, j\in s_t} \frac{\Delta_{ij}}{\pi_{ij}} \frac{1}{\hat{p}_{ij}^{1\to t}} \frac{u_{it}}{\pi_i} \frac{u_{jt}}{\pi_j}. \tag{3.11}$$

Similarly, replacing in (2.12) the variable y_{it} with u_{it} yields the estimator of the variance due to the non-response

$$\hat{V}_t^{\mathrm{nr}}\left(\hat{\theta}_t\right) = \sum_{\delta=1}^{t} \sum_{i\in s_t} \frac{\hat{p}_i^{\delta}\left(1-\hat{p}_i^{\delta}\right)}{\hat{p}_i^{\delta\to t}} \left(\frac{u_{it}}{\pi_i \hat{p}_i^{1\to\delta}} - k_i^{\delta}\left(\hat{h}_i^{\delta}\right)^{\top} \hat{\gamma}_{t\theta}^{\delta}\right)^2 \tag{3.12}$$

$$\hat{\gamma}_{t\theta}^{\delta} = \left\{\sum_{i\in s_t} k_i^{\delta} \frac{\hat{p}_i^{\delta}\left(1-\hat{p}_i^{\delta}\right)}{\hat{p}_i^{\delta\to t}} \hat{h}_i^{\delta}\left(\hat{h}_i^{\delta}\right)^{\top}\right\}^{-1} \sum_{i\in s_t} \frac{1-\hat{p}_i^{\delta}}{\hat{p}_i^{1\to t}}\, \hat{h}_i^{\delta}\, \frac{u_{it}}{\pi_i}. \tag{3.13}$$

The global variance estimator for $\hat{\theta}_t$ is

$$\hat{V}_t\left(\hat{\theta}_t\right) = \hat{V}_t^p\left(\hat{\theta}_t\right) + \hat{V}_t^{\mathrm{nr}}\left(\hat{\theta}_t\right). \tag{3.14}$$

The simplified estimator of the variance due to non-response is

$$\hat{V}_{t,\,\mathrm{simp}}^{\mathrm{nr}}\left(\hat{\theta}_t\right) = \sum_{i\in s_t} \frac{1-\hat{p}_i^{1\to t}}{\left(\hat{p}_i^{1\to t}\right)^2} \left(\frac{u_{it}}{\pi_i}\right)^2. \tag{3.15}$$

The bias of this simplified variance estimator will depend on the explanatory power for $\hat{h}_i^{\delta}$ on the linearized variable u_{it}.

3.3 Variance estimation for complex parameters under calibration

The calibrated weights w_{ti} may be used to obtain an estimator of the parameter $\theta(t)$. Substituting $\hat{Y}_{wt}$ into $\theta(t) = f\{Y(t)\}$ yields the calibrated plug-in estimator $\hat{\theta}_{wt} = f\left(\hat{Y}_{wt}\right)$. To obtain a variance estimator for $\hat{\theta}_{wt}$, we first compute the estimated linearized variable $u_{it} = \left\{f'\left(\hat{Y}_t\right)\right\}^{\top} y_{it}$ and take

$$e_{\theta it} = u_{it} - \hat{b}_{\theta t} x_i \tag{3.16}$$

with

$$\hat{b}_{\theta t} = \left(\sum_{i\in s_t} \frac{1}{\pi_i \hat{p}_i^{1\to t}} x_i x_i^{\top}\right)^{-1} \sum_{i\in s_t} \frac{1}{\pi_i \hat{p}_i^{1\to t}} x_i u_{it}. \tag{3.17}$$

Replacing in (2.11) the variable y_{it} with $e_{\theta it}$ yields the estimator of the variance due to the sampling design

$$\hat{V}_t^p\left(\hat{\theta}_{wt}\right) = \sum_{i,\, j\in s_t} \frac{\Delta_{ij}}{\pi_{ij}} \frac{1}{\hat{p}_{ij}^{1\to t}} \frac{e_{\theta it}}{\pi_i} \frac{e_{\theta jt}}{\pi_j}. \tag{3.18}$$

Similarly, replacing in (2.12) the variable y_{it} with $e_{\theta it}$ yields the estimator of the variance due to the non-response

$$\hat{V}_t^{\mathrm{nr}}\left(\hat{\theta}_{wt}\right) = \sum_{\delta=1}^{t}\sum_{i\in s_t}\frac{\hat{p}_i^{\delta}\left(1-\hat{p}_i^{\delta}\right)}{\hat{p}_i^{\delta\to t}}\left(\frac{e_{\theta it}}{\pi_i\hat{p}_i^{1\to\delta}} - k_i^{\delta}\left(\hat{h}_i^{\delta}\right)^{\top}\hat{\gamma}_{te\theta}^{\delta}\right)^2 \tag{3.19}$$

$$\hat{\gamma}_{te\theta}^{\delta} = \left\{\sum_{i\in s_t} k_i^{\delta}\frac{\hat{p}_i^{\delta}\left(1-\hat{p}_i^{\delta}\right)}{\hat{p}_i^{\delta\to t}}\hat{h}_i^{\delta}\left(\hat{h}_i^{\delta}\right)^{\top}\right\}^{-1}\sum_{i\in s_t}\frac{1-\hat{p}_i^{\delta}}{\hat{p}_i^{1\to t}}\hat{h}_i^{\delta}\frac{e_{\theta it}}{\pi_i}. \tag{3.20}$$

The global variance estimator for $\hat{\theta}_{wt}$ is

$$\hat{V}_t\left(\hat{\theta}_{wt}\right) = \hat{V}_t^{p}\left(\hat{\theta}_{wt}\right) + \hat{V}_t^{\mathrm{nr}}\left(\hat{\theta}_{wt}\right). \tag{3.21}$$

The simplified estimator of the variance due to non-response is

$$\hat{V}_{t,\,\mathrm{simp}}^{\mathrm{nr}}\left(\hat{\theta}_{wt}\right) = \sum_{i\in s_t}\frac{1-\hat{p}_i^{1\to t}}{\left(\hat{p}_i^{1\to t}\right)^2}\left(\frac{e_{\theta it}}{\pi_i}\right)^2. \tag{3.22}$$

Since the variable $e_{\theta it}$ is obtained as the residual in the regression of the linearized variable u_{it} on the calibration variables x_i, the explanatory power for $\hat{h}_i^{\delta}$ on $e_{\theta it}$ is expected to be small in practice, and the bias of the simplified variance estimator is expected to be small as well.

4 Longitudinal estimators

We may be interested in a change in parameters, such as

$$\Delta(u\to t) = Y(t) - Y(u), \tag{4.1}$$

the difference between the totals of a variable of interest measured at two different times $u < t$. Since the variable y_{iu} is measured on all sub-samples $s_{u'}$ for $u' = u,\ldots,t$, there are several possible estimators for $\Delta(u\to t)$. For $u' = u,\ldots,t$, we denote by

$$\hat{\Delta}_{u't}(u\to t) = \sum_{i\in s_t}\frac{y_{it}}{\pi_i\hat{p}_i^{1\to t}} - \sum_{i\in s_{u'}}\frac{y_{iu}}{\pi_i\hat{p}_i^{1\to u'}} \tag{4.2}$$

the estimator which makes use of s_t for the estimation of $Y(t)$, and of $s_{u'}$ for the estimation of $Y(u)$. The case $u' = u$ corresponds to the estimation of $Y(u)$ on the largest available sub-sample, s_u. The case $u' = t$ corresponds to the estimation of $Y(u)$ and $Y(t)$ on the common sub-sample s_t.

In the context of full response, several authors have recommended the estimator $\hat{\Delta}_{tt}(u\to t)$ which makes use of the common sample only, if the variables y_{ui} and y_{ti} are strongly positively correlated; see Caron and Ravalet (2000), Qualité and Tillé (2008), Goga, Deville and Ruiz-Gazen (2009), Chauvet and Goga (2018). In our context, this choice may be heuristically justified as follows. For $u' < t$, and by conditioning on the sub-sample $s_{u'}$, we obtain

$$V\left\{\hat{\Delta}_{u't}(u\to t)\right\} \simeq V\left\{\sum_{i\in s_{u'}}\frac{y_{it}-y_{iu}}{\pi_i\hat{p}_i^{1\to u'}}\right\} + EV\left\{\sum_{i\in s_t}\frac{y_{it}}{\pi_i\hat{p}_i^{1\to t}}\,\middle|\, s_{u'}\right\}, \tag{4.3}$$

$$V\left\{\hat{\Delta}_{tt}(u \to t)\right\} \simeq V\left\{\sum_{i \in s_{u'}} \frac{y_{it} - y_{iu}}{\pi_i \hat{p}_i^{1 \to u'}}\right\} + EV\left\{\sum_{i \in s_t} \frac{y_{it} - y_{iu}}{\pi_i \hat{p}_i^{1 \to t}} \,\middle|\, s_{u'}\right\}. \tag{4.4}$$

In equations (4.3) and (4.4), the first term in the right-hand side is identical. Since the variables y_{iu} and y_{it} are expected to be positively correlated, the difference $y_{it} - y_{iu}$ is expected to be smaller than y_{it}. Therefore, the estimator $\hat{\Delta}_{tt}(u \to t)$ based on the common sample is expected to be more efficient in terms of variance. The results of a small simulation study in Section 5.2 support this heuristic reasoning. Therefore, we focus only in this Section on the estimator $\hat{\Delta}_{tt}(u \to t)$ for the estimation of $\Delta(u \to t)$. As pointed out by a Referee, and following the approach in Zhou and Kim (2012), we may obtain a gain in efficiency by using the full information on s_u, namely by calibrating the weights $\left(\pi_i \hat{p}_i^{1 \to t}\right)^{-1}$ on the estimator $\hat{Y}_u$.

Replacing in (2.11) the variable y_{it} with $y_{it} - y_{iu}$ yields the estimator of the variance due to the sampling design

$$\hat{V}_t^p\left\{\hat{\Delta}_{tt}(u \to t)\right\} = \sum_{i,\, j \in s_t} \frac{\Delta_{ij}}{\pi_{ij}} \frac{1}{\hat{p}_{ij}^{1 \to t}} \frac{(y_{it} - y_{iu})}{\pi_i} \frac{\left(y_{jt} - y_{ju}\right)}{\pi_j}. \tag{4.5}$$

Similarly, replacing in (2.12) the variable y_{it} with $y_{it} - y_{iu}$ yields the estimator of the variance due to the non-response

$$\hat{V}_t^{\text{nr}}\left\{\hat{\Delta}_{tt}(u \to t)\right\} = \sum_{\delta=1}^{t} \sum_{i \in s_t} \frac{\hat{p}_i^\delta\left(1 - \hat{p}_i^\delta\right)}{\hat{p}_i^{\delta \to t}} \left(\frac{y_{it} - y_{iu}}{\pi_i \hat{p}_i^{1 \to \delta}} - k_i^\delta \left(\hat{h}_i^\delta\right)^\top \hat{\gamma}_{t\Delta}^\delta\right)^2 \tag{4.6}$$

with

$$\hat{\gamma}_{t\Delta}^\delta = \left\{\sum_{i \in s_t} k_i^\delta \frac{\hat{p}_i^\delta\left(1 - \hat{p}_i^\delta\right)}{\hat{p}_i^{\delta \to t}} \hat{h}_i^\delta \left(\hat{h}_i^\delta\right)^\top\right\}^{-1} \sum_{i \in s_t} \frac{1 - \hat{p}_i^\delta}{\hat{p}_i^{1 \to t}} \hat{h}_i^\delta \frac{y_{it} - y_{iu}}{\pi_i}. \tag{4.7}$$

The global variance estimator for $\hat{\Delta}_{tt}(u \to t)$ is

$$\hat{V}_t\left\{\hat{\Delta}_{tt}(u \to t)\right\} = \hat{V}_t^p\left\{\hat{\Delta}_{tt}(u \to t)\right\} + \hat{V}_t^{\text{nr}}\left\{\hat{\Delta}_{tt}(u \to t)\right\}. \tag{4.8}$$

Variance estimation for measures of change is also considered in Berger (2004), Qualité and Tillé (2008), Goga et al. (2009), Chauvet and Goga (2018), among others.

The simplified estimator of the variance due to non-response is

$$\hat{V}_{t,\,\text{simp}}^{\text{nr}}\left\{\hat{\Delta}_{tt}(u \to t)\right\} = \sum_{i \in s_t} \frac{1 - \hat{p}_i^{1 \to t}}{\left(\hat{p}_i^{1 \to t}\right)^2} \left(\frac{y_{it} - y_{iu}}{\pi_i}\right)^2. \tag{4.9}$$

If the variables y_{it} and y_{iu} are strongly positively correlated, the bias of the simplified variance estimator is expected to be small.

5 A simulation study

In this section, several artificial populations are generated according to the model described in Section 5.1. In Section 5.2, we consider several estimators for a change between totals, which illustrates the heuristic reasoning in Section 4. A Monte Carlo experiment is presented in Section 5.3, and several variance estimators for estimating a total, a ratio or a parameter change are compared. The results from Tables 5.1 and 5.2 are readily reproducible using the R code provided in the Supplementary Material.

5.1 Simulation set-up

We consider seven populations of size 10,000, each containing three variables of interest y_{i1}, y_{i2} and y_{i3} observed at times $t = 1, 2$ and 3, respectively. The variables of interest are generated according to the superpopulation model

$$y_{i1} = \alpha^0 + \alpha^a x_{ai} + \alpha^b x_{bi} + \sigma u_{i1}, \tag{5.1}$$

$$y_{i2} = \rho y_{i1} + \sigma u_{i2}, \tag{5.2}$$

$$y_{i3} = \rho y_{i2} + \sigma u_{i3}. \tag{5.3}$$

The auxiliary variables x_{ai} and x_{bi} are independently generated from a Gamma distribution with shape and scale parameters 2 and 1. Two auxiliary variables x_{ci} and x_{di}, not related to the variables of interest, are generated similarly. The variables u_{i1}, u_{i2} and u_{i3} are independently generated according to a standard normal distribution. We use $\alpha^0 = 10, \alpha^a = \alpha^b = 5$ and $\sigma = 10,$ which leads to a coefficient of determination (R^2) in model (5.1) approximately equal to 0.50. The parameter ρ is set to 0, 0.2, 0.4, 0.6, 0.8, 1.0 and 1.2 for populations 1 to 7, respectively.

For each population, a simple random sample s_0 of size $n = 1{,}000$ is selected. Three non-response phases are then successively simulated. At each phase $\delta = 1, 2, 3,$ the sub-sample of respondents s_δ is obtained by Poisson sampling with a response probability p_i^δ for unit $i,$ defined as

$$\text{logit}\left(p_i^\delta\right) = \beta^{\delta 0} + \beta^{\delta a} x_{ai} + \beta^{\delta b} x_{bi}. \tag{5.4}$$

We use $\beta^{\delta 0} = -1$ at each phase $\delta = 1, 2, 3.$ For $\delta = 1,$ we use $\beta^{1a} = \beta^{1b} = 0.60,$ which corresponds to an average response rate of 0.75. For $\delta = 2, 3,$ we use $\beta^{\delta a} = \beta^{\delta b} = 0.75,$ which corresponds to an average response rate of 0.81. Inside each sub-sample $s_\delta,$ the estimated response probabilities $\hat{p}_i^\delta$ are obtained by means of an unweighted logistic regression.

5.2 Comparison of estimators for a difference of totals

In this section, we are interested in comparing the accuracy of two estimators for a difference of totals $\Delta(u \to t)$ for $u = 1$ and $t = 2,$ for $u = 1$ and $t = 3,$ and for $u = 2$ and $t = 3.$ We consider the estimator $\hat{\Delta}_{ut}(u \to t),$ which makes use of the whole appropriate sub-samples for variables y_{iu} and $y_{it},$

and the estimator $\hat{\Delta}_{tt}(u \to t)$, which makes use of the common sub-sample only. These two estimators are compared through the relative difference (RD) of their variances, which are defined as follows:

$$\text{RD}(u \to t) = 100 \times \frac{V\left\{\hat{\Delta}_{ut}(u \to t)\right\} - V\left\{\hat{\Delta}_{tt}(u \to t)\right\}}{V\left\{\hat{\Delta}_{tt}(u \to t)\right\}}. \tag{5.5}$$

The true variances are replaced by their Monte Carlo approximation, obtained by repeating $B = 100{,}000$ times the sample selection and the non-response phases.

The results are presented in Table 5.1. A positive RD indicates that the use of the common sample only leads to a more accurate estimator. As could be expected, the RD increases in all cases with ρ, that is, when the correlation between y_{it} and y_{iu} increases. For $u = 1$ and $t = 2$, and for $u = 2$ and $t = 3$, the estimator $\hat{\Delta}_{tt}(u \to t)$ is more accurate for ρ greater than 0.6. For $u = 1$ and $t = 3$, $\hat{\Delta}_{tt}(u \to t)$ is more accurate for ρ greater than 0.8.

Table 5.1
Relative Difference (RD) between two estimators for a difference of totals

ρ	$\mathbf{RD}(1 \to 2)$	$\mathbf{RD}(1 \to 3)$	$\mathbf{RD}(2 \to 3)$
0.0	-12	-27	-13
0.2	-09	-25	-11
0.4	-04	-20	-03
0.6	05	-09	11
0.8	17	11	39
1.0	30	33	83
1.2	40	46	127

5.3 Performances of the variance estimators

In this section, we consider the artificial population 5 $(\rho = 0.8)$ generated as described in Section 5.1. The sample selection by means of simple random sampling of size $n = 1{,}000$ and the three non-response phases are applied $B = 5{,}000$ times. We are interested in evaluating the variance estimators and the simplified variance estimators, in case of estimating a total, a ratio or a change in totals.

As for the total $Y(t)$, we consider at each time $t = 1, 2, 3,$ three estimators. The estimator $\hat{Y}_t$ makes use of the weights $d_{ti} = \pi_i^{-1}\left(\hat{p}_i^{1 \to t}\right)^{-1}$. The estimator $\hat{Y}_{wt}$ makes use of the weights w_i, obtained by calibrating the weights d_{ti} on the population size and on the totals of the auxiliary variables x_{ai} and x_{bi}. The estimator $\hat{Y}_{\tilde{w}t}$ makes use of the weights $\tilde{w}_i$, obtained by calibrating the weights d_{ti} on the population size and on the totals of the auxiliary variables x_{ci} and x_{di}. The working model is therefore well-specified for $\hat{Y}_{wt}$, but not for $\hat{Y}_{\tilde{w}t}$. The proposed variance estimator for $\hat{Y}_t$ is obtained from equation (2.16), and the simplified variance estimator is obtained by plugging in (2.16) the simplified variance estimator for non-response given in (2.17). The proposed variance estimators for $\hat{Y}_{wt}$ and $\hat{Y}_{\tilde{w}t}$ are obtained from equation (3.8), and the simplified variance estimators are obtained by plugging in (3.8) the simplified variance estimator for non-response given in (3.9).

We are also interested in estimating the ratio $R(t) = Y(t)/Y(1)$ for $t = 2, 3.$ At each time $t,$ we consider three estimators. The estimator $\hat{R}_t$ makes use of the weights $d_i.$ The proposed variance estimator is obtained from equation (3.14), by using the estimated linearized variable $u_{it} = \left(\hat{Y}_1\right)^{-1}\left(y_{ti} - \hat{R}_t y_{1i}\right).$ The simplified variance estimator is obtained by plugging in (3.14) the simplified variance estimator for non-response given in (3.15). The estimators $\hat{R}_{wt}$ and $\hat{R}_{\tilde{w}t}$ make use of the calibrated weights w_i and $\tilde{w}_i.$ The proposed variance estimators are obtained from equation (3.21). The simplified variance estimators are obtained by plugging in (3.21) the simplified variance estimator for non-response given in (3.22).

Finally, we are interested in estimating the change in totals $\Delta(1 \rightarrow t)$ for $t = 2, 3.$ At each time $t,$ we consider three estimators. The estimator $\hat{\Delta}_{tt}(1 \rightarrow t)$ makes use of the weights $d_i.$ The proposed variance estimator is obtained from equation (4.8), and the simplified variance estimator is obtained by plugging in (4.8) the simplified variance estimator for non-response given in (4.9). The estimators $\hat{\Delta}_{tt,w}(1 \rightarrow t)$ and $\hat{\Delta}_{tt,\tilde{w}}(1 \rightarrow t)$ make use of the calibrated weights w_i and $\tilde{w}_i.$ The proposed variance estimators are obtained from equation (4.8), by replacing $y_{it} - y_{iu}$ by the estimated residual for the weighted regression of $y_{it} - y_{iu}$ on the calibration variables. The simplified variance estimators are obtained by plugging in (4.8) the simplified variance estimator for non-response given in (4.9).

For a proposed variance estimator $\hat{V},$ we computed the Monte Carlo Percent Relative Bias

$$\text{RB}_{\text{mc}}\left(\hat{V}\right) = 100 \times \frac{B^{-1}\sum_{b=1}^{B}\hat{V}^{(b)} - V}{V}$$

where the global variance V was approximated through an independent set of 100,000 simulations. To evaluate the contribution of some component $\hat{V}_a$ into the variance estimator $\hat{V},$ we computed the contribution (in percent)

$$\text{CONTR}_{\text{mc}}\left(\hat{V}_a\right) = 100 \times \frac{\frac{1}{B}\sum_{b=1}^{B}\hat{V}_a^{(b)}}{\frac{1}{B}\sum_{b=1}^{B}\hat{V}^{(b)}}.$$

To evaluate the simplified variance estimator for the non-response $\hat{V}_{\text{simp}}^{\text{nr}},$ we computed the Monte Carlo Percent Relative Bias

$$\text{RB}_{\text{mc}}\left(\hat{V}_{\text{simp}}^{\text{nr}}\right) = 100 \times \frac{B^{-1}\sum_{b=1}^{B}\hat{V}_{\text{simp}}^{(b)} - V^{\text{nr}}}{V^{\text{nr}}},$$

where the variance V^{nr} due to non-response was approximated through an independent set of 100,000 simulations.

The simulation results are presented in Table 5.2. The proposed variance estimator is almost unbiased in all cases. As could be expected, the contribution of the variance due to the sampling design decreases with time, as the number of respondents decreases and as the variance due to non-response becomes larger. The simplified variance estimator is highly biased for the variance due to non-response in case of $\hat{Y}_t.$ The bias decreases quickly with time, but remains large at time $t = 3.$ The simplified variance estimator is almost unbiased for a calibrated estimator when the working model is adequately specified, but is severely biased otherwise. This is consistent with our reasoning in Section 3.1. The simplified variance estimator is almost unbiased for the three estimators of the ratio, and for the calibrated estimators of the change in totals. In case of the non-calibrated estimator for the change in totals, the bias can be as high as 30%.

Table 5.2
Relative bias of a global variance estimator, relative contribution to the estimators of variance components and relative bias of a simplified variance estimator for the variance due to non-response for the estimation of a total, a ratio or a change in totals with three sets of weights

	$t = 1$	$t = 2$	$t = 3$	$t = 1$	$t = 2$	$t = 3$	$t = 1$	$t = 2$	$t = 3$
		$\hat{Y}_t$			$\hat{Y}_{wt}$			$\hat{Y}_{\tilde{w}t}$	
$\text{RB}_{\text{mc}}(\hat{V})$	0	-1	-2	-1	-1	-2	-1	-1	-3
$\text{CONTR}_{\text{mc}}(\hat{V}_t^{p})$	81	57	35	69	49	32	80	56	35
$\text{CONTR}_{\text{mc}}(\hat{V}_t^{\text{nr}1})$	19	19	13	31	22	15	20	18	13
$\text{CONTR}_{\text{mc}}(\hat{V}_t^{\text{nr}2})$	-	25	18	-	28	19	-	25	17
$\text{CONTR}_{\text{mc}}(\hat{V}_t^{\text{nr}3})$	-	-	34	-	-	34	-	-	34
$\text{RB}_{\text{mc}}(\hat{V}_{t,\text{simp}}^{\text{nr}})$	559	188	80	0	-1	-2	83	34	15
		$\hat{R}_t$			$\hat{R}_{wt}$			$\hat{R}_{\tilde{w}t}$	
$\text{RB}_{\text{mc}}(\hat{V})$	-	0	-2	-	-1	-2	-	-1	-2
$\text{CONTR}_{\text{mc}}(\hat{V}_t^{p})$	-	49	32	-	49	32	-	50	33
$\text{CONTR}_{\text{mc}}(\hat{V}_t^{\text{nr}1})$	-	22	15	-	22	15	-	22	15
$\text{CONTR}_{\text{mc}}(\hat{V}_t^{\text{nr}2})$	-	28	19	-	28	19	-	28	19
$\text{CONTR}_{\text{mc}}(\hat{V}_t^{\text{nr}3})$	-	-	34	-	-	34	-	-	34
$\text{RB}_{\text{mc}}(\hat{V}_{t,\text{simp}}^{\text{nr}})$	-	0	0	-	-1	-2	-	-1	-1
		$\hat{\Delta}_{tt}(1 \to t)$			$\hat{\Delta}_{tt,w}(1 \to t)$			$\hat{\Delta}_{tt,\tilde{w}}(1 \to t)$	
$\text{RB}_{\text{mc}}(\hat{V})$	-	0	-2	-	0	-2	-	-1	-3
$\text{CONTR}_{\text{mc}}(\hat{V}_t^{p})$	-	50	33	-	49	32	-	50	33
$\text{CONTR}_{\text{mc}}(\hat{V}_t^{\text{nr}1})$	-	22	14	-	22	15	-	22	14
$\text{CONTR}_{\text{mc}}(\hat{V}_t^{\text{nr}2})$	-	28	18	-	28	19	-	28	18
$\text{CONTR}_{\text{mc}}(\hat{V}_t^{\text{nr}3})$	-	-	34	-	-	34	-	-	34
$\text{RB}_{\text{mc}}(\hat{V}_{t,\text{simp}}^{\text{nr}})$	-	19	30	-	-1	-2	-	3	5

6 Illustration

In this section, we aim at illustrating our results on a real data set from the ELFE survey. The population of inference consists of infants born in one of the 544 French maternity units during 2011, except very premature infants. Our illustration is meant to mimic as closely as possible the methodology of the ELFE survey. In particular, the modeling of attrition at each time is performed with variables available at baseline as explanatory variables only. As pointed out by the Associate Editor, under the MAR assumption, the variables of interest measured at any times $\delta < t$ may also have been used to model attrition between times $t - 1$ and t.

An original sample s_0 of about 35,600 infants was originally selected when the babies were just a few days old and were still at the maternity unit. The sample was selected using a cross-classified sampling design (Skinner, 2015; Juillard, Chauvet and Ruiz-Gazen, 2016). A sample of days and a sample of maternity units were independently selected, and both sample selections may be approximated by stratified simple random sampling (STSI). The sample consisted in all the infants born during one of the 25 selected days in one of the 320 selected maternity units.

Among the 35,600 infants originally selected, a total of 18,329 face-to-face interviews were completed with their families, which represents a response rate of 51%. This led to the subsample s_1 after accounting for non-response. The weights at time $t = 1$ were computed on the basis of the original sampling weights, adjusted in two steps. First, response probabilities were estimated by means of a model of Response Homogeneity Groups (RHGs), with 20 RHGs defined by using a logistic regression model with explanatory variables *Age of the mother*, *Gemellary identity* and *Season of birth*. Then, a calibration by means of the raking ratio method was performed on the binary variables *Born within marriage, Immigrant mother* and *Gemellary identity*.

When the children reached the age of two months, the parents had the first phone interview with a response rate of 87%. This leads to the subsample s_2. The weights at time $t = 2$ were computed on the basis on the weight obtained at time $t = 1$, with a two-step adjustment. First, response probabilities were estimated by means of 20 RHGs, defined by using a logistic regression with explanatory variables *Age of the mother*, *Mother nationality* and *Father present at childbirth*. Then, a calibration by the raking ratio method was performed on the same calibration variables as at time $t = 1$.

When the children were one year old, the parents were contacted by phone with a response rate of 77%. This led to the subsample s_3. The weights at time $t = 3$ were computed on the basis on the weights obtained at time $t = 2$, with a two-step adjustment similar to that realized at time $t = 2$.

We considered three variables of interest: *Breastfeeding exclusivity at the childbirth, at two month, at one year*. For each of these variables, we computed the estimator $\hat{R}_t$ and the calibrated estimator $\hat{R}_{wt}$ for the percentage $R(t)$ of breastfeeding among all the children at time t, and the associated variance estimators. We also computed the estimated coefficient of variation (in percent), defined as

$$\widehat{\mathrm{CV}}_t\left(\hat{Y}_t\right) = 100 \times \frac{\sqrt{\hat{V}_t\left(\hat{Y}_t\right)}}{\hat{Y}_t}. \tag{6.1}$$

For each component $\hat{V}_{ta}$ in the estimated variance $\hat{V}_t$, we computed its contribution (in percent) defined as

$$\text{CONTR}\left(\hat{V}_{ta}\right) = 100 \times \frac{\hat{V}_{ta} - \hat{V}_t}{\hat{V}_t}. \tag{6.2}$$

We also computed the simplified variance estimator for non-response $\hat{V}^{\text{nr}}_{t,\,\text{simp}}$, and the relative difference (in percent) with the approximately unbiased variance estimator $\hat{V}^{\text{nr}}$ defined as

$$\text{RD}\left(\hat{V}^{\text{nr}}_{t,\,\text{simp}}\right) = 100 \times \frac{\hat{V}^{\text{nr}}_{t,\,\text{simp}} - \hat{V}^{\text{nr}}_t}{\hat{V}^{\text{nr}}_t}. \tag{6.3}$$

The results are given in Table 6.1. As observed in the simulation study, the RD of the simplified variance estimator for non-response is negligible in all cases.

Table 6.1
Estimates for a ratio, variance estimates, coefficient of variation, relative contributions of variance components and relative difference of a simplified variance estimator for a variable in the ELFE survey

Breastfeeding exclusivity	***t* = 1 maternity**	***t* = 2 2 months**	***t* = 3 1 year**	***t* = 1 maternity**	***t* = 2 2 months**	***t* = 3 1 year**
	without calibration			**with calibration**		
$\hat{R}_t$ (%)	59.0	30.6	3.3	59.4	31.0	3.4
$\hat{V}\left(\hat{R}_t\right)$	1.34E-05	1.50E-05	2.58E-06	1.28E-05	1.48E-05	2.60E-06
$\hat{\text{CV}}\left(\hat{Y}_t\right)$(%)	0.6	1.3	4.8	0.6	1.2	4.7
$\text{CONTR}\left(\hat{V}_t^{p}\right)$	31	34	24	28	34	25
$\text{CONTR}\left(\hat{V}_t^{\text{nr}1}\right)$	69	51	42	72	51	41
$\text{CONTR}\left(\hat{V}_t^{\text{nr}2}\right)$	-	15	13	-	15	13
$\text{CONTR}\left(\hat{V}_t^{\text{nr}3}\right)$	-	-	21	-	-	21
$\text{RD}\left(\hat{V}^{\text{nr}}_{t,\,\text{simp}}\right)$	2	2	0	1	2	0

7 Conclusion

In this paper, we considered variance estimation accounting for weighting adjustments in panel surveys. We proposed both an approximately unbiased variance estimator and a simplified variance estimator for estimators of totals, complex parameters and measures of change, which covers most cases that may be encountered in practice. Our simulation results indicate that the proposed variance estimator performs well in all cases considered. The simplified variance estimator tends to overestimate the variance of the expansion estimator for totals, and to overestimate the variance for calibrated estimators of totals when the calibration variables lack of explanatory power for the variable of interest. However, the simplified variance estimator performs well for the estimation of ratios and change in totals with calibrated weights, even if the calibration model is not appropriate for the study variable.

The assumption of independent response behaviour is usually not tenable for multi-stage surveys, since units within clusters tend to be correlated with respect to the response behaviour. In this context, estimation of response probabilities based upon conditional logistic regression in the context of correlated responses has been studied by Skinner and D'Arrigo (2011), see also Kim, Kwon and Park (2016). Extending the present work in the context of correlated response behaviour is a challenging problem for further research.

Acknowledgements

We thank the Editors, an Associate Editor and the referees for useful comments and suggestions which led to an improvement of the paper.

Appendix

Estimation of the variance due to non-response for Response Homogeneity Groups

We consider the model of Response Homogeneity Groups introduced in Section 2.5. Recall that this model may be summarized as follows: at each time $\delta = 1, \ldots, t,$ the sub-sample $s_{\delta-1}$ is partitioned into $C(\delta - 1)$ groups $s_{\delta-1}^{c}, c = 1, \ldots, C(\delta - 1).$ The response probabilities are assumed to be constant within the groups.

This model is equivalent to the logistic regression model in (2.18), with

$$z_i^{\delta} = \left[1\left\{i \in s_{\delta-1}^{1}\right\}, \ldots, 1\left\{i \in s_{\delta-1}^{C(\delta-1)}\right\}\right]^{\top}. \tag{A.1}$$

The equation (2.2) leads to the estimated response probabilities

$$\hat{p}_i^{\delta} = \frac{\sum_{i \in s_{\delta-1}^{c}} k_i^{\delta} r_i^{\delta}}{\sum_{i \in s_{\delta-1}^{c}} k_i^{\delta}} \qquad \text{for} \qquad i \in s_{\delta-1}^{c}. \tag{A.2}$$

We first consider the case when the reweighted estimator is computed at time $t = 1.$ In the estimator of the variance due to non-response given in (2.21), the vector $\hat{\gamma}_1^1$ simplifies as

$$\hat{\gamma}_1^1 = \left(\frac{\sum_{i \in s_1 \cap s_0^1} \frac{y_{i1}}{\pi_i}}{\hat{p}_1^1 \sum_{i \in s_1 \cap s_0^1} k_i^1}, \ldots, \frac{\sum_{i \in s_1 \cap s_0^{C(0)}} \frac{y_{i1}}{\pi_i}}{\hat{p}_{C(0)}^1 \sum_{i \in s_1 \cap s_0^{C(0)}} k_i^1}\right)^{\top}. \tag{A.3}$$

After some algebra, the variance estimator in (2.21) may be rewritten as

$$\hat{V}_1^{\mathrm{nr}}\left(\hat{Y}_1\right) = \sum_{c=1}^{C(0)} \frac{\left(1 - \hat{p}_c^1\right)}{\left(\hat{p}_c^1\right)^2} \sum_{i \in s_1 \cap s_0^c} \left(\frac{y_{i1}}{\pi_i} - k_i^1 \frac{\sum_{j \in s_1 \cap s_0^c} \frac{y_{j1}}{\pi_j}}{\sum_{j \in s_1 \cap s_0^c} k_j^1}\right)^2. \tag{A.4}$$

We now consider the case when the reweighted estimator is computed at time $t = 2$. We focus on the simpler case when the same system of RHGs is kept over time. In the estimator of the variance due to non-response given in (2.22), the vectors $\hat{\gamma}_2^1$ and $\hat{\gamma}_2^2$ simplify as

$$\hat{\gamma}_2^1 = \left(\frac{\sum_{i \in s_2 \cap s_1^1} \frac{y_{i2}}{\pi_i}}{\hat{p}_1^1 \sum_{i \in s_2 \cap s_1^1} k_i^1}, \ldots, \frac{\sum_{i \in s_2 \cap s_1^{C(0)}} \frac{y_{i2}}{\pi_i}}{\hat{p}_{C(0)}^1 \sum_{i \in s_2 \cap s_1^{C(0)}} k_i^1} \right)^\top, \tag{A.5}$$

$$\hat{\gamma}_2^2 = \left(\frac{\sum_{i \in s_2 \cap s_1^1} \frac{y_{i2}}{\pi_i}}{\hat{p}_1^1 \hat{p}_1^2 \sum_{i \in s_2 \cap s_1^1} k_i^2}, \ldots, \frac{\sum_{i \in s_2 \cap s_1^{C(0)}} \frac{y_{i2}}{\pi_i}}{\hat{p}_{C(0)}^1 \hat{p}_{C(0)}^2 \sum_{i \in s_2 \cap s_1^{C(0)}} k_i^2} \right)^\top. \tag{A.6}$$

After some algebra, the variance estimator in (2.22) may be rewritten as

$$\begin{aligned} \hat{V}_2^{\text{nr}}\left(\hat{Y}_2\right) = & \sum_{c=1}^{C(0)} \frac{\left(1 - \hat{p}_c^1\right)}{\hat{p}_c^2} \sum_{i \in s_2 \cap s_1^c} \left(\frac{y_{i2}}{\pi_i \hat{p}_c^1} - k_i^1 \frac{\sum_{j \in s_2 \cap s_1^c} \frac{y_{j2}}{\pi_j}}{\sum_{j \in s_2 \cap s_1^c} k_j^1} \right)^2 \\ & + \sum_{c=1}^{C(0)} \left(1 - \hat{p}_c^2\right) \sum_{i \in s_2 \cap s_1^c} \left(\frac{y_{i2}}{\pi_i \hat{p}_c^1 \hat{p}_c^2} - k_i^2 \frac{\sum_{j \in s_2 \cap s_1^c} \frac{y_{j2}}{\pi_j}}{\sum_{j \in s_2 \cap s_1^c} k_j^2} \right)^2. \end{aligned} \tag{A.7}$$

If we further assume that k_i^δ is constant over times $\delta = 1, 2,$ and may thus be rewritten as k_i, the expression in (A.7) simplifies as

$$\hat{V}_2^{\text{nr}}\left(\hat{Y}_2\right) = \sum_{c=1}^{C(0)} \frac{\left(1 - \hat{p}_c^{1\rightarrow 2}\right)}{\left(\hat{p}_c^{1\rightarrow 2}\right)^2} \sum_{i \in s_2 \cap s_1^c} \left(\frac{y_{i2}}{\pi_i} - k_i \frac{\sum_{j \in s_2 \cap s_1^c} \frac{y_{j2}}{\pi_j}}{\sum_{j \in s_2 \cap s_1^c} k_j} \right)^2. \tag{A.8}$$

with $\hat{p}_c^{1\rightarrow 2} = \prod_{\delta=1}^{2} \hat{p}_c^\delta$ for $c = 1, \ldots, C(0)$. This simplification of the variance estimator can be extended to the reweighted estimator at time t. Assuming that the RHGs are kept over time, and that $k_i^\delta = k_i$ for any $\delta = 1, \ldots, t,$ the variance estimator in (2.12) may be written as

$$\hat{V}_t^{\text{nr}}\left(\hat{Y}_t\right) = \sum_{c=1}^{C(0)} \frac{\left(1 - \hat{p}_c^{1\rightarrow t}\right)}{\left(\hat{p}_c^{1\rightarrow t}\right)^2} \sum_{i \in s_t \cap s_{t-1}^c} \left(\frac{y_{it}}{\pi_i} - k_i \frac{\sum_{j \in s_t \cap s_{t-1}^c} \frac{y_{jt}}{\pi_j}}{\sum_{j \in s_t \cap s_{t-1}^c} k_j} \right)^2 \tag{A.9}$$

with $\hat{p}_c^{1\rightarrow t} = \prod_{\delta=1}^{t} \hat{p}_c^\delta$ for $c = 1, \ldots, C(0)$.

References

Beaumont, J.-F. (2005). Calibrated imputation in surveys under a quasimodel-assisted approach. *Journal of the Royal Statistical Society, Series B*, 67, 445-458.

Beaumont, J.-F., and Haziza, D. (2016). A note on the concept of invariance in two-phase sampling designs. *Survey Methodology*, 42, 2, 319-323. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2016002/article/14662-eng.pdf.

Berger, Y. (2004). Variance estimation for measures of change in probability sampling. *Canadian Journal of Statistics*, 32, 4, 451-467.

Caron, N., and Ravalet, P. (2000). Estimation dans les enquêtes répétées : application à l'enquête emploi en continu. Technical report INSEE, Paris.

Chauvet, G., and Goga, C. (2018). Linearization versus bootstrap for variance estimation of the change between Gini indexes. *Survey Methodology*, 44, 1, 17-42. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2018001/article/54926-eng.pdf.

Clarke, P., and Tate, P. (2002). An application of non-ignorable non-response models for gross flows estimation in the British labour force survey. *Australian & New Zealand Journal of Statistics*, 4, 413-425.

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

Ekholm, A., and Laaksonen, S. (1991). Weighting via response modeling in the finnish household budget survey. *Journal of Official Statistics*, 7, 325-327.

Fay, R. (1992). When are inferences from multiple imputation valid? *Proceedings of the Survey Research Methods Section*, American Statistical Association, 81, 1, 227-232.

Fuller, W., and An, A. (1998). Regression adjustment for non-response. *Journal of the Indian Society of Agricultural Statistics*, 51, 331-342.

Fuller, W.A., Loughin, M.M. and Baker, H.D. (1994). Regression weighting in the presence of nonresponse with application to the 1987-1988 Nationwide Food Consumption Survey. *Survey Methodology*, 20, 1, 75-85. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/1994001/article/14429-eng.pdf.

Goga, C., Deville, J.-C. and Ruiz-Gazen, A. (2009). Composite estimation and linearization method for two-sample survey data. *Biometrika*, 96, 691-709.

Hawkes, D., and Plewis, I. (2009). Modelling nonresponse in the national child development study. *Journal of the Royal Statistical Society, Series A*, 169, 479-491.

Juillard, H., Chauvet, G. and Ruiz-Gazen, A. (2017). Estimation under cross-classified sampling with application to a childhood survey. *Journal of the American Statistical Association*, 112, 850-858.

Kalton, G. (2009). Design for surveys over time. *Handbook of Statistics*, 29, 89-108.

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., Kwon, Y. and Park, M. (2016). Calibrated propensity score method for survey nonresponse in cluster sampling. *Biometrika*, 103, 461-473.

Laaksonen, S. (2007). Weighting for two-phase surveyed data. *Survey Methodology*, 33, 2, 121-130. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2007002/article/10489-eng.pdf.

Laaksonen, S., and Chambers, R.L. (2006). Survey estimation under informative nonresponse with follow-up. *Journal of Official Statistics*, 22, 81-95.

Laniel, N. (1988). Variances for a rotating sample from a changing population. *Proceedings of the Business and Economics Statistics Section*, American Statistical Association, 246-250.

Laurie, H., Smith, R. and Scott, L. (1999). Strategies for reducing nonresponse in a longitudinal panel survey. *Journal of Official Statistics*, 15, 269-282.

Lynn, P. (2009). Methods for longitudinal surveys. *Methodology of Longitudinal Surveys*, 1-19.

Nordberg, L. (2000). On variance estimation for measures of change when samples are coordinated by the use of permanent random numbers. *Journal of Official Statistics*, 16, 363-378.

Pirus, C., Bois, C., Dufourg, M., Lanoë, J., Vandentorren, S., Leridon, H. and the Elfe team (2010). Constructing a cohort: Experience with the French Elfe project. *Population*, 65, 637-670.

Qualité, L., and Tillé, Y. (2008). Variance estimation of changes in repeated surveys and its application to the Swiss survey of value added. *Survey Methodology*, 34, 2, 173-181. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2008002/article/10758-eng.pdf.

Rendtel, U., and Harms, T. (2009). Weighting and calibration for household panels. *Methodology of Longitudinal Surveys*, 265-286.

Rizzo, L., Kalton, G. and Brick, J.M. (1996). A comparison of some weighting adjustment methods for panel nonresponse. *Survey Methodology*, 22, 1, 43-53. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/1996001/article/14386-eng.pdf.

Silva, P., and Skinner, C. (1997). Cross-classiffed sampling: Some estimation theory. *Variable Selection for Regression Estimation in Finite Populations*, 23, 23-32.

Skinner, C. (2015). Cross-classiffed sampling: Some estimation theory. *Statistics & Probability Letters*, 104, 163-168.

Skinner, C., and D'Arrigo, J. (2011). Inverse probability weighting for clustered non-response. *Biometrika*, 98, 953-966.

Skinner, C., and Vieira, M. (2005). Design effects in the analysis of longitudinal survey data. S3RI Methdology Working Papers, M05/13. Southampton, UK: Southampton Statistical Sciences Research Institute.

Slud, E.V., and Bailey, L. (2010). Evaluation and selection of models for attrition nonresponse adjustment. *Journal of Official Statistics*, 26, 1-18.

Tam, S. (1984). On covariance from overlapping samples. *The American Statistician*, 38, 1-18.

Vandecasteele, L., and Debels, A. (2007). Attrition in panel data: The effectiveness of weighting. *European Sociological Review*, 23, 1, 81-97.

Zhou, M., and Kim, J. (2012). An effcient method of estimation for longitudinal surveys with monotone missing data. *Biometrika*, 99, 631-648.

How to decompose the non-response variance: A total survey error approach

Keven Bosa, Serge Godbout, Fraser Mills and Frédéric Picard[1]

Abstract

When a linear imputation method is used to correct non-response based on certain assumptions, total variance can be assigned to non-responding units. Linear imputation is not as limited as it seems, given that the most common methods – ratio, donor, mean and auxiliary value imputation – are all linear imputation methods. We will discuss the inference framework and the unit-level decomposition of variance due to non-response. Simulation results will also be presented. This decomposition can be used to prioritize non-response follow-up or manual corrections, or simply to guide data analysis.

Key Words: Total variance; Adaptive design; Imputation.

1 Introduction

Total survey error is described by Biemer (2010) as the “accumulation of all errors that may arise in the design, collection, processing and analysis of survey data”. He classified survey error components into sampling error and nonsampling errors, such as, non-response, coverage, measurement and data processing errors. These errors may affect variance, bias, or both. The total survey error paradigm aims at maximizing survey quality by minimizing total survey error within prespecified resource constraints like budget, people, or time.

At Statistics Canada, the Corporate Business Architecture initiated the Integrated Business Statistics Program (IBSP) as the standardized platform for more than 140 economic surveys with the objective of achieving efficiency, enhancing quality and improving responsiveness. In particular, reducing collection costs while managing non-response error was identified as one of the program’s pillars. Consequently, an adaptive design where different units may receive different treatments became a keystone for this program. For more details on IBSP, see Statistics Canada (2015). Groves and Heeringa (2006) showed how paradata could be used to increase the response rate. Schouten, Calinescu and Luiten (2013) gave a general framework for an adaptive design and explained how the R-indicator could be used in this context.

A new survey process model called Rolling Estimates has been developed as an attempt to address the IBSP’s pillar mentioned above. The Rolling Estimates model is based on iterative processing and estimation cycles throughout the collection period. Basically, the idea of this model is to compute key estimates with their associated quality indicators at several specific times during the collection period. At the beginning, all units are assigned to the self-response survey treatment which means that the respondents are asked to complete the online questionnaire. Collection efforts like computer-assisted telephone interview non-response follow-ups are then performed on units contributing the most to the estimates where the quality is low based on the preliminary results of the Rolling Estimates. This can be viewed as an adaptive design since the treatments on the units depend on the quality of the estimates produced during the collection period. Most of the work

[1] Keven Bosa, Serge Godbout, Fraser Mills and Frédéric Picard, Statistics Canada, 100 Tunney’s Pasture Driveway, Ottawa, Ontario, K1A 0T6. E-mail: serge.godbout@canada.ca.

regarding the development of the IBSP's adaptive design has been done since 2010. Godbout, Beaucage and Turmelle (2011), Turmelle, Godbout and Bosa (2012), Mills, Godbout, Bosa and Turmelle (2013) and Bosa and Godbout (2014) made use of this idea in the context of the IBSP adaptive design to minimize the number of follow-ups in order to reach a targeted quality in terms of coefficient of variation.

This paper revisits the work done so far for IBSP and presents an approach to decompose non-response variance into an item-level score for a given variable of interest within a domain. This item score is basically an attempt to estimate the contribution to the variance borrowed by a single unit. Units with a large score will contribute the most to reduce the variance and the coefficient of variation which is often used as a quality indicator in surveys. However, there are generally many important variables and domains in a survey. The proposed approach first computes, for a given unit, item-level scores for important variables and domains. Then, item scores can be combined into a single unit-level score in order to rank units. For example, the unit score can be a weighted sum or the maximum of its item scores. The most attractive use of the resulting unit-level score is to prioritize units, the ones with the highest scores, for the most expensive collection operations such as telephone follow-up, computer-assisted telephone interview or computer-assisted personal interview. This paper assumes total and partial non-response are both treated in the adaptive design, but treatments may be different depending on the type of non-response. For instance, telephone follow-ups could be made in the case of total non-response whereas questionnaires with partial non-response could be reviewed by analysts. This type of adaptive design generates strong interactions between collection operations, observed data and measured quality. Bosa and Godbout (2014) showed how this methodology was implemented in IBSP under the Rolling Estimates model.

Emphasis will be placed on the derivation of the item-level score throughout this paper. Therefore, the special case of only one variable of interest within a domain is studied. Also, only one imputation method is used to impute the variable of interest in the case of non-response so as to simplify the notation and to ease comprehension for the reader.

Section 2 describes the inference framework. In Section 3, the decomposition of the variance at the unit-level is expressed. In other words, the contribution of each nonresponding unit to the variance is computed. A simulation study was conducted to evaluate the proposed score. It is described in Section 4. Finally, Section 5 expresses some thoughts and conclusions.

2 Inference framework

Assume a sample s of size n is drawn from a population U of size N. Define the population total by

$$t_d = \sum_{k \in U} d_k y_k \tag{2.1}$$

for a variable, y, and a domain indicator, d_k, which takes the value $d_k = 1$ if unit k belongs to the domain d, and $d_k = 0$ otherwise. In the context of full response, t_d is estimated by $\hat{t}_d^0 = \sum_{k \in s} d_k w_k y_k$ where w_k could be the sampling weight or a calibrated weight if calibration is performed. Because surveys are generally subject to non-response, both unit or item, a sample unit is classified into either a responding or a nonresponding unit with regard to the variable y at any given point during data collection. The subset s_r contains item-responding units whereas s_m contains item-nonresponding units. Note that s_r and s_m, respectively of size n_r and n_m, form a partition of the sample s, $P_s = \{s_r, s_m\}$, with $s_r \cup s_m = s$ and $s_r \cap s_m = \varnothing$.

The approach proposed in this paper assumes that imputation is used in case of non-response, which is the common approach in business surveys. Moreover, this approach can be considered for both item and unit non-response as long as imputation is used. However, since only one variable of interest y is considered here for simplicity, then no distinction is made if the y variable is imputed because of item or unit non-response. Also, the set s_r and s_m are not indexed by an item number for simplicity without loss of generality. However, the action following the calculation of a unit score might be different depending on whether the unit is responding or not.

2.1 Estimation under imputation

The framework requires linear imputation methods. In other words, the imputed value, y_k^*, can be written as a linear combination of the values reported by the other units. This linear combination is given by $y_k^* = \varphi_{0k} + \sum_{l \in s_r} \varphi_{lk} y_l$. The quantities, φ_{0k} and φ_{lk} do not depend on the values of variable of interest, y, but they may depend on s, s_r and auxiliary data from the nonrespondents available on the frame, registers or elsewhere. Linear imputation methods cover most methods used in practice like auxiliary value imputation (Beaumont, Haziza and Bocci, 2011) and linear regression imputation, as well as donor imputation, which is often used to impute categorical variables.

It is common practice to use several imputation methods, referred to as composite imputation, applied sequentially to the same variable. More than one linear imputation method can be used to impute nonresponding units. Section 2 of Beaumont and Bissonnette (2011) defines composite imputation in detail. Briefly, suppose that the set of nonrespondents is broken down into two or more groups and that a different imputation method is used within each group. For example, let $\mathbf{x}_k$ be the complete vector of auxiliary variables for unit k, and suppose regression imputation is used to impute the variable of interest. However, if, for some cases, $\mathbf{x}_k$ were incomplete, another imputation method, based on the available subset of $\mathbf{x}_k$, would be used. The approach presented in our paper can be generalized to include composite imputation as long as linear imputation methods are used. For simplicity of notation, the case of a single linear imputation method is presented.

The estimator of the domain total after imputation is given by

$$\hat{t}_d = \sum_{l \in s_r} w_l d_l y_l + \sum_{k \in s_m} w_k d_k y_k^* \tag{2.2}$$

where w_k is the sampling weight or a calibrated weight. The estimator presented in equation (2.2) can be rewritten as

$$\begin{aligned}
\hat{t}_d &= \sum_{l \in s_r} w_l d_l y_l + \sum_{k \in s_m} w_k d_k y_k^* \\
&= \sum_{l \in s_r} w_l d_l y_l + \sum_{k \in s_m} w_k d_k \left(\varphi_{0k} + \sum_{l \in s_r} \varphi_{lk} y_l \right) \\
&= \sum_{l \in s_r} w_l d_l y_l + \sum_{k \in s_m} w_k d_k \varphi_{0k} + \sum_{l \in s_r} y_l \sum_{k \in s_m} w_k d_k \varphi_{lk} \\
&= W_{0d} + \sum_{l \in s_r} w_l d_l y_l + \sum_{l \in s_r} y_l W_{dl} \\
&= W_{0d} + \sum_{l \in s_r} y_l \left(w_l d_l + W_{dl} \right).
\end{aligned}$$

The quantities W_{dl} and W_{0d} denote the compensatory weights (or adjustment weights) defined as

$$\begin{aligned} W_{dl} &= \sum_{k \in s_m} w_k d_k \varphi_{lk} \\ W_{0d} &= \sum_{k \in s_m} w_k d_k \varphi_{0k}. \end{aligned}$$

They represent the effect of the non-response in the domain, d, carried by the respondent unit, $l \in s_r$, with a reported value, y_l.

2.2 Variance estimation

Consider an imputation model, η, describing the relationship between variable y and the vector of observed auxiliary variables $\mathbf{x}^{\text{obs}}$. Let $E_\eta(.)$, $\mathrm{Var}_\eta(.)$ and $\mathrm{cov}_\eta(.)$ denote respectively the expectation, the variance, and the covariance with respect to the imputation model η. The imputation model is

$$\begin{aligned} E_\eta\left(y_k \mid \mathbf{X}^{\text{obs}}\right) &= \mu_k \\ V_\eta\left(y_k \mid \mathbf{X}^{\text{obs}}\right) &= \sigma_k^2 \\ \mathrm{cov}_\eta\left(y_k, y_{k'} \mid \mathbf{X}^{\text{obs}}\right) &= 0 \end{aligned}$$

where $k, k' \in U$ and $k \neq k'$. The matrix $\mathbf{X}^{\text{obs}}$ contains all observed vectors $\mathbf{x}^{\text{obs}}$. The quantities μ_k and σ_k^2 can be estimated by $\hat{\mu}_k$ and $\hat{\sigma}_k^2$ respectively. We assume that these estimators are unbiased with respect to the imputation model η. These estimators will be useful later for estimating the total variance components and the unit decompositions of those components.

The total error of the estimator (2.2) can be expressed as

$$\hat{t}_d - t_d = \left(\hat{t}_d^0 - t_d\right) + \left(\hat{t}_d - \hat{t}_d^0\right), \tag{2.3}$$

where $\hat{t}_d^0$ is the estimator under complete response given by (2.1). The first term on the right-hand side of (2.3) is usually referred to as the sampling error and the second term is called the non-response error. As proposed in Särndal (1992) and in Beaumont and Bissonnette (2011), the mean square error of $\hat{t}_d$ using (2.3) can be decomposed in three components and is given by

$$\begin{aligned} E_{\eta pq}\left(\hat{t}_d - t_d\right)^2 = {} & E_\eta V_p\left(\hat{t}_d\right) + E_{pq}E_\eta\left[\left(\hat{t}_d - \hat{t}_d^0\right)^2 \mid s, s_r\right] \\ & + 2E_{pq}E_\eta\left[\left(\hat{t}_d - \hat{t}_d^0\right)\left(\hat{t}_d^0 - t_d\right) \mid s, s_r\right], \end{aligned} \tag{2.4}$$

under imputation model, η, sampling design, p, and response mechanism, q. $E_{\eta pq}\left(\hat{t}_d - t_d\right)^2$ is approximately equivalent to the variance $V_{\eta pq}\left(\hat{t}_d - t_d\right)$ assuming that the overall bias is negligible. Thus, the equation (2.4) is equivalent to $V_{\eta pq}\left(\hat{t}_d - t_d\right) \equiv V_{\text{TOT}}\left(\hat{t}_d\right) = V_{\text{SAM}}\left(\hat{t}_d\right) + V_{\text{NR}}\left(\hat{t}_d\right) + V_{\text{MIX}}\left(\hat{t}_d\right)$, where:

- $V_{\text{SAM}}\left(\hat{t}_d\right) \equiv E_\eta V_p\left(\hat{t}_d\right)$ is the sampling variance;
- $V_{\text{NR}}\left(\hat{t}_d\right) \equiv E_{pq}E_\eta\left[\left(\hat{t}_d - \hat{t}_d^0\right)^2 \mid s, s_r\right]$ is the non-response variance;
- $V_{\text{MIX}}\left(\hat{t}_d\right) \equiv 2E_{pq}E_\eta\left[\left(\hat{t}_d - \hat{t}_d^0\right)\left(\hat{t}_d^0 - t_d\right) \mid s, s_r\right]$ is the covariance between sampling and non-response error terms, also called the mixed variance component.

Beaumont and Bissonnette (2011) proposed the following estimators for $V_{\text{SAM}}\left(\hat{t}_d\right)$, $V_{\text{NR}}\left(\hat{t}_d\right)$ and $V_{\text{MIX}}\left(\hat{t}_d\right)$.

1. $\hat{V}_{\text{SAM}}\left(\hat{t}_d\right) = \hat{V}_{\text{ORD}}\left(\hat{t}_d\right) + \hat{V}_{\text{DIF}}\left(\hat{t}_d\right)$ where:
   - $\hat{V}_{\text{ORD}}\left(\hat{t}_d\right)$ is the naive sampling variance estimator using the imputed values as though they were reported values.
   - $\hat{V}_{\text{DIF}}\left(\hat{t}_d\right) = \sum_{k \in s_m} (1 - \pi_k)\, w_k^2 d_k \hat{\sigma}_k^2$ is a correction to $\hat{V}_{\text{ORD}}\left(\hat{t}_d\right)$ in order to reduce the bias of $\hat{V}_{\text{ORD}}\left(\hat{t}_d\right)$, as proposed by Beaumont and Bocci (2009), since the variance component $\hat{V}_{\text{ORD}}\left(\hat{t}_d\right)$ relies on the use of imputed values, usually more homogeneous than the reported values.
2. $\hat{V}_{\text{NR}}\left(\hat{t}_d\right) = \sum_{l \in s_r} W_{dl}^2 \hat{\sigma}_l^2 + \sum_{k \in s_m} w_k^2 d_k \hat{\sigma}_k^2$ is the estimator of the non-response component of variance.
3. $\hat{V}_{\text{MIX}}\left(\hat{t}_d\right) = 2\sum_{l \in s_r} W_{dl}\,(w_l - 1)\, d_l \hat{\sigma}_l^2 - 2\sum_{k \in s_m} w_k\,(w_k - 1)\, d_k \hat{\sigma}_k^2$ is the estimator of the mixed variance component.

Under complete response, $s_m = \varnothing$, the compensation weights are $W_{dl} = 0$, and the variance components, $\hat{V}_{\text{DIF}}\left(\hat{t}_d\right), \hat{V}_{\text{NR}}\left(\hat{t}_d\right)$, and $\hat{V}_{\text{MIX}}\left(\hat{t}_d\right)$, are also equal to 0, leaving the total variance as $\hat{V}_{\text{TOT}}\left(\hat{t}_d\right) = \hat{V}_{\text{ORD}}\left(\hat{t}_d\right)$. Under a census, $s = U$, the variance components, $\hat{V}_{\text{DIF}}\left(\hat{t}_d\right), \hat{V}_{\text{ORD}}\left(\hat{t}_d\right)$, and $\hat{V}_{\text{MIX}}\left(\hat{t}_d\right)$, are equal to 0, leaving the total variance as $\hat{V}_{\text{TOT}}\left(\hat{t}_d\right) = \hat{V}_{\text{NR}}\left(\hat{t}_d\right)$.

2.3 Non-response bias

The reduction of non-response bias is always a desirable goal. It can be achieved through an adaptive design and/or through an appropriate method of dealing with missing values. Our framework assumes that the non-response bias is removed through imputation methods that use relevant auxiliary information. In practice, it is likely that imputation will only reduce non-response bias, not eliminate it. We may then wonder whether adaptive designs could be used to reduce further the bias. In the context of non-response weighting, Beaumont, Bocci and Haziza (2014) argued that auxiliary information used in an adaptive design to reduce non-response bias can also be used in non-response weighting to reduce the same amount of bias. Their argument can also be made in the context of imputation. This justifies our focus on variance reduction rather than bias reduction. We acknowledge that some bias may remain after imputation but ignore this bias because it may not be possible to reduce it further through an adaptive design without the availability of additional auxiliary information. However, it is possible to reduce the variance through an adaptive design.

3 Unit-level error decomposition of variance components

This section describes the approach used to evaluate the contribution of a given nonresponding unit, $\lambda \in s_m$, to the estimated total variance for the estimation of a total for a given variable.

The unit-level error decomposition, δ_λ, of the total variance for a given unit, λ, is defined as the difference between the estimated total variance, and the projected total variance, i.e., $\delta_\lambda\left(\hat{V}_{\text{TOT}}\left(\hat{t}_d\right)\right) \equiv \hat{V}_{\text{TOT}}\left(\hat{t}_d\right) - \hat{V}_{\text{TOT}}^{(\lambda)}\left(\hat{t}_d\right)$. The superscript (λ) is used to indicate projected quantities when unit λ is converted to a respondent. So, $\delta_{\lambda}\left(\hat{V}_{\text{TOT}}\left(\hat{t}_d\right)\right)$ can be seen as the expected gain, in terms of total variance, of converting a nonrespondent unit λ to a respondent.

In order to get $\delta_{\lambda}\left(\hat{V}_{\text{TOT}}\left(\hat{t}_d\right)\right)$, λ is moved from s_m to s_r, generating the new partition $P_s^{(\lambda)}$ of the sample from P_s where $P_s^{(\lambda)} = \left\{s_r^{(\lambda)}, s_m^{(\lambda)}\right\}$, $s_r^{(\lambda)} = s_r \cup \{\lambda\}$ and $s_m^{(\lambda)} = s_m \setminus \{\lambda\}$, as illustrated in Figure 3.1.

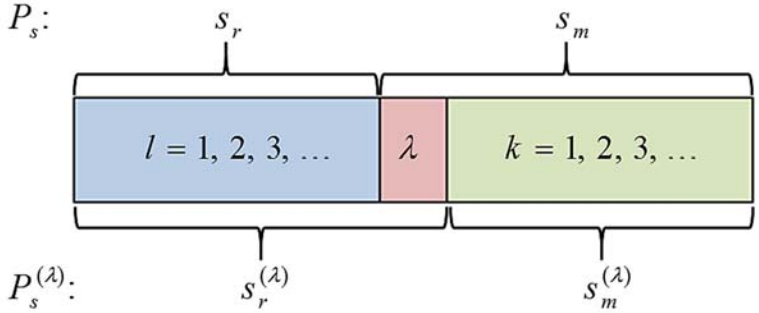


Figure 3.1 Sample partitions.

Some assumptions are necessary to decompose the variance components. It is recognized that these assumptions may not perfectly hold in reality. However, they can be used to generate accurate results, as shown in the simulation in Section 4. The required assumptions are:

1. Projected reported value: let $\lambda \in s_m$ be converted to a response and let $y_{\lambda}^{(\lambda)} = y_{\lambda}^{*}$.
2. Projected imputation parameters: $\forall k \in s_m$, $\hat{\mu}_k^{(\lambda)} = \hat{\mu}_k$ and $\hat{\sigma}_k^{(\lambda)} = \hat{\sigma}_k$.
3. Projected imputation relationship matrix: $\forall k \in s_m$ and $\forall l \in s_r$, $\varphi_{lk}^{(\lambda)} = 0$ if $l = \lambda$ or if $k = \lambda$ or $\varphi_{lk}^{(\lambda)} = \varphi_{lk}$ otherwise. Similarly, $\varphi_{0k}^{(\lambda)} = 0$ if $k = \lambda$ or $\varphi_{0k}^{(\lambda)} = \varphi_{0k}$ otherwise.

Assumption 1 implies that if a nonresponding unit, λ, would have been converted to a respondent, its reported value is equal to its imputed value. This is not true generally, but the imputed value is our best estimate. The expectation is that this imputed value is close enough to the reported value to estimate the error on the variance components. This assumption will have an impact when the sampling variance is decomposed.

Assumption 2 states that the estimated parameters of the imputation model would remain unchanged if λ were a respondent. In the case of a consistent imputation model parameter estimator, this assumption becomes more realistic when s_r is larger.

Finally, assumption 3 means that the imputation relationship between nonrespondents and respondents remains unchanged, except when unit λ is involved. In other words, the converted unit, λ, is no longer imputed from respondents, but will not be used to impute other nonresponding units. Figure 3.2 shows how assumption 3 is reflected in terms of the phi matrix.

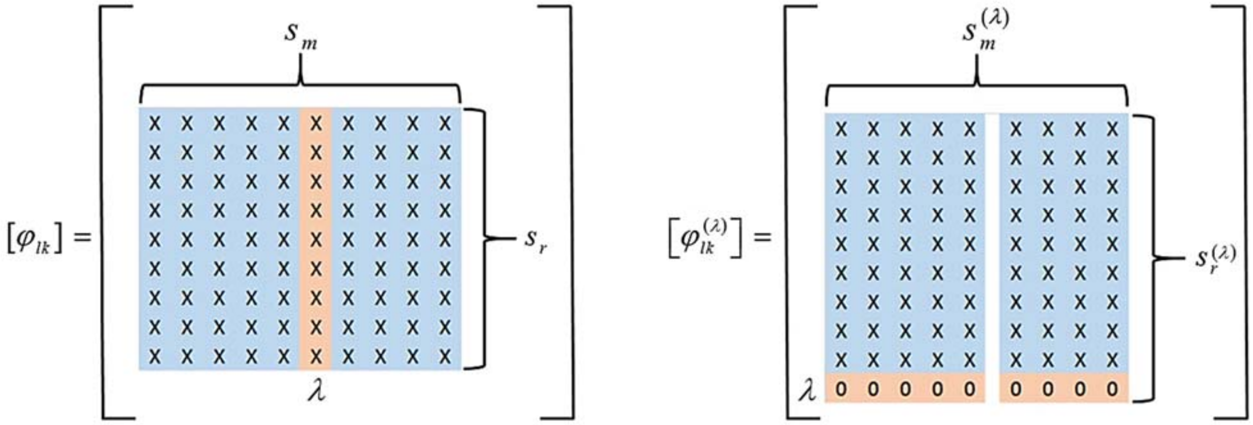


Figure 3.2 Initial and projected imputation relationship phi matrix.

Therefore, the compensation weight, $W_{dl}^{(\lambda)},$ of a responding unit, $\forall l \in s_r,$ is projected as

$$\begin{aligned} W_{dl}^{(\lambda)} &= \sum_{k \in s_m^{(\lambda)}} w_k d_k \varphi_{lk}^{(\lambda)} \\ &= \sum_{k \in s_m} w_k d_k \varphi_{lk} - w_\lambda d_\lambda \varphi_{l\lambda} \\ &= W_{dl} - w_\lambda d_\lambda \varphi_{l\lambda}. \end{aligned} \tag{3.1}$$

The marginal weight from the converted unit λ is withdrawn from the original compensation weight, $W_{dl},$ to obtain the new $W_{dl}^{(\lambda)}.$ Note that $W_{d\lambda}^{(\lambda)} = \sum_{k \in s_m^{(\lambda)}} w_k d_k \varphi_{\lambda k}^{(\lambda)} = 0$ because $\varphi_{\lambda k}^{(\lambda)} = 0$ under assumption 3. As mentioned above, it means that λ isn't used to impute nonrespondents.

In the next subsections, the unit-level error decomposition for unit λ is computed for the four variance components, as described in Section 2.3.

3.1 Unit-level error decomposition of the naive sampling variance

The quantity $\hat{V}_{\text{ORD}}\left(\hat{t}_d\right)$ depends on the y – values, the final weights and the first-order and second-order selection probabilities. The unit-level error decomposition of the naive sampling variance component $\hat{V}_{\text{ORD}}\left(\hat{t}_d\right)$ is trivial since the assumption that unit λ goes from s_m to s_r does not change weights and selection probabilities. Under assumption 1, the projected reported value $y_\lambda^{(\lambda)}$ is set to y_λ^* so that $\hat{V}_{\text{ORD}}^{(\lambda)}\left(\hat{t}_d\right) = \hat{V}_{\text{ORD}}\left(\hat{t}_d\right)$ when λ is converted to a responding unit. Consequently, the decomposition of $\hat{V}_{\text{ORD}}\left(\hat{t}_d\right)$ is given by

$$\delta_\lambda\left(\hat{V}_{\text{ORD}}\left(\hat{t}_d\right)\right) \equiv \hat{V}_{\text{ORD}}\left(\hat{t}_d\right) - \hat{V}_{\text{ORD}}^{(\lambda)}\left(\hat{t}_d\right) = 0. \tag{3.2}$$

This result is consistent with the idea that the naive sampling variance point estimate will likely change, but it is not expected to decrease with an extra responding unit.

3.2 Unit-level decomposition of the correction to the sampling variance component

The unit-level error decomposition for unit λ of the correction to the sampling variance component, $\hat{V}_{\text{DIF}}\left(\hat{t}_d\right),$ is given by

$$\begin{aligned}\delta_\lambda\left(\hat{V}_{\text{DIF}}\left(\hat{t}_d\right)\right) &\equiv \hat{V}_{\text{DIF}}\left(\hat{t}_d\right) - \hat{V}_{\text{DIF}}^{(\lambda)}\left(\hat{t}_d\right)\\ &= \sum_{k\in s_m}\left(1-\pi_k\right)d_k w_k^2\hat{\sigma}_k^2 - \sum_{\lambda\in s_m^{(\lambda)}}\left(1-\pi_k\right)d_k w_k^2\left(\hat{\sigma}_k^{(\lambda)}\right)^2.\end{aligned}$$

Under assumption 2, $\hat{\sigma}_k^{(\lambda)} = \hat{\sigma}_k,$ so that

$$\delta_\lambda\left(\hat{V}_{\text{DIF}}\left(\hat{t}_d\right)\right) = \left(1-\pi_\lambda\right)d_\lambda w_\lambda^2\hat{\sigma}_\lambda^2. \tag{3.3}$$

The astute reader will notice that the actual sampling variance (not its estimation) should not be impacted by whether or not a unit is a respondent. However, we decided to include the impact of a unit on the sampling variance estimation in order to be coherent in the way we treat the three components $V_{\text{SAM}}\left(\hat{t}_d\right), V_{\text{NR}}\left(\hat{t}_d\right)$ and $V_{\text{MIX}}\left(\hat{t}_d\right).$

3.3 Unit-level decomposition of the non-response variance component

The unit-level error decomposition for unit λ of the non-response variance component $\hat{V}_{\text{NR}}\left(\hat{t}_d\right)$ is given by

$$\begin{aligned}\delta_\lambda\left(\hat{V}_{\text{NR}}\left(\hat{t}_d\right)\right) &\equiv \hat{V}_{\text{NR}}\left(\hat{t}_d\right) - \hat{V}_{\text{NR}}^{(\lambda)}\left(\hat{t}_d\right)\\ &= \left(\sum_{l\in s_r}W_{dl}^2\hat{\sigma}_l^2 + \sum_{k\in s_m}w_k^2 d_k\hat{\sigma}_k^2\right) - \left(\sum_{l\in s_r^{(\lambda)}}\left(W_{dl}^{(\lambda)}\right)^2\left(\hat{\sigma}_l^{(\lambda)}\right)^2 + \sum_{k\in s_m^{(\lambda)}}w_k^2 d_k\left(\hat{\sigma}_k^{(\lambda)}\right)^2\right).\end{aligned}$$

Under assumptions 2 and 3, $\hat{\sigma}_k^{(\lambda)} = \hat{\sigma}_k$ and $W_{d\lambda}^{(\lambda)} = 0.$ This can be rewritten as

$$\delta_\lambda\left(\hat{V}_{\text{NR}}\left(\hat{t}_d\right)\right) = \left(\sum_{l\in s_r}W_{dl}^2\hat{\sigma}_l^2 - \sum_{l\in s_r}\left(W_{dl}^{(\lambda)}\right)^2\hat{\sigma}_l^2\right) + w_\lambda^2 d_\lambda\hat{\sigma}_\lambda^2.$$

Using formula (3.1), this becomes

$$\begin{aligned}\delta_\lambda\left(\hat{V}_{\text{NR}}\left(\hat{t}_d\right)\right) &= \left(\sum_{l\in s_r}W_{dl}^2\hat{\sigma}_l^2 - \sum_{l\in s_r}\left(W_{dl} - w_\lambda d_\lambda\varphi_{l\lambda}\right)^2\hat{\sigma}_l^2\right) + w_\lambda^2 d_\lambda\hat{\sigma}_\lambda^2\\ &= \left(\sum_{l\in s_r}W_{dl}^2\hat{\sigma}_l^2 - \left(W_{dl}^2 - 2W_{dl}w_\lambda d_\lambda\varphi_{l\lambda} + w_\lambda^2 d_\lambda\varphi_{l\lambda}^2\right)\hat{\sigma}_l^2\right) + w_\lambda^2 d_\lambda\hat{\sigma}_\lambda^2\\ &= \sum_{l\in s_r}\left(2W_{dl}w_\lambda d_\lambda\varphi_{l\lambda} - w_\lambda^2 d_\lambda\varphi_{l\lambda}^2\right)\hat{\sigma}_l^2 + w_\lambda^2 d_\lambda\hat{\sigma}_\lambda^2. \qquad (3.4)\end{aligned}$$

3.4 Unit-level decomposition of the mixed variance component

Finally, the impact of unit λ on the variance component term, $\hat{V}_{\text{MIX}}\left(\hat{t}_d\right)$, is given by

$$\begin{aligned}\delta_\lambda\left(\hat{V}_{\text{MIX}}\left(\hat{t}_d\right)\right) &\equiv \hat{V}_{\text{MIX}}\left(\hat{t}_d\right) - \hat{V}_{\text{MIX}}^{(\lambda)}\left(\hat{t}_d\right)\\ &= \left(2\sum_{l\in s_r} W_{dl}\left(w_l - 1\right) d_l \hat{\sigma}_l^2 - 2\sum_{k\in s_m} w_k\left(w_k - 1\right) d_k \hat{\sigma}_k^2\right)\\ &\quad - \left(2\sum_{l\in s_r^{(\lambda)}} W_{dl}^{(\lambda)}\left(w_l - 1\right) d_l \left(\hat{\sigma}_l^{(\lambda)}\right)^2 - 2\sum_{k\in s_m^{(\lambda)}} w_k\left(w_k - 1\right) d_k \left(\hat{\sigma}_k^{(\lambda)}\right)^2\right).\end{aligned}$$

This equation can be rewritten as follows, under assumptions 2 and 3 and equation (3.1)

$$\begin{aligned}\delta_\lambda\left(\hat{V}_{\text{MIX}}\left(\hat{t}_d\right)\right) &= \left(2\sum_{l\in s_r} W_{dl}\left(w_l - 1\right) d_l \hat{\sigma}_l^2 - 2\sum_{k\in s_m} w_k\left(w_k - 1\right) d_k \hat{\sigma}_k^2\right)\\ &\quad - \left(2\sum_{l\in s_r}\left(W_{dl} - w_\lambda d_\lambda \varphi_{l\lambda}\right)\left(w_l - 1\right) d_l \hat{\sigma}_l^2 - 2\sum_{k\in s_m^{(\lambda)}} w_k\left(w_k - 1\right) d_k \hat{\sigma}_k^2\right)\\ &= 2\sum_{l\in s_r} w_\lambda d_\lambda \varphi_{l\lambda}\left(w_l - 1\right) d_l \hat{\sigma}_l^2 - 2 w_\lambda\left(w_\lambda - 1\right) d_\lambda \hat{\sigma}_\lambda^2. \qquad (3.5)\end{aligned}$$

In Section 2.3, the estimation of the total variance, $\hat{V}_{\text{TOT}}\left(\hat{t}_d\right)$, has been defined as $\hat{V}_{\text{TOT}}\left(\hat{t}_d\right) = \hat{V}_{\text{ORD}}\left(\hat{t}_d\right) + \hat{V}_{\text{DIF}}\left(\hat{t}_d\right) + \hat{V}_{\text{NR}}\left(\hat{t}_d\right) + \hat{V}_{\text{MIX}}\left(\hat{t}_d\right)$. Similarly, the impact of unit λ on $\hat{V}_{\text{TOT}}\left(\hat{t}_d\right)$ is defined as

$$\delta_\lambda\left(\hat{V}_{\text{TOT}}\left(\hat{t}_d\right)\right) = \delta_\lambda\left(\hat{V}_{\text{ORD}}\left(\hat{t}_d\right)\right) + \delta_\lambda\left(\hat{V}_{\text{DIFF}}\left(\hat{t}_d\right)\right) + \delta_\lambda\left(\hat{V}_{\text{NR}}\left(\hat{t}_d\right)\right) + \delta_\lambda\left(\hat{V}_{\text{MIX}}\left(\hat{t}_d\right)\right),$$

where $\delta_\lambda\left(\hat{V}_{\text{ORD}}\left(\hat{t}_d\right)\right)$, $\delta_\lambda\left(\hat{V}_{\text{DIF}}\left(\hat{t}_d\right)\right)$, $\delta_\lambda\left(\hat{V}_{\text{NR}}\left(\hat{t}_d\right)\right)$, and $\delta_\lambda\left(\hat{V}_{\text{MIX}}\left(\hat{t}_d\right)\right)$ are respectively given by equations (3.2), (3.3), (3.4) and (3.5).

It can be observed (proofs are given in the appendix) that $\hat{V}_{\text{DIF}}\left(\hat{t}_d\right) = \sum_{k\in s_m} \delta_k\left(\hat{V}_{\text{DIF}}\left(\hat{t}_d\right)\right)$ and $\hat{V}_{\text{MIX}}\left(\hat{t}_d\right) = \sum_{k\in s_m} \delta_k\left(\hat{V}_{\text{MIX}}\left(\hat{t}_d\right)\right)$. However, this linear relation doesn't hold for $\hat{V}_{\text{NR}}\left(\hat{t}_d\right)$. This property is important to consider because, for $\hat{V}_{\text{DIF}}\left(\hat{t}_d\right)$ and $\hat{V}_{\text{MIX}}\left(\hat{t}_d\right)$, the sum of the unit-level errors on all nonresponding units, $k \in s_m$, is equal to the corresponding estimated variance component. In the case of non-response variance component, the sum of the errors is different than $\hat{V}_{\text{NR}}\left(\hat{t}_d\right)$. The difference is given by

$$\sum_{k\in s_m} \delta_k\left(\hat{V}_{\text{NR}}\left(\hat{t}_d\right)\right) - \hat{V}_{\text{NR}}\left(\hat{t}_d\right) = \sum_{l\in s_r}\left(\left(\sum_{k\in s_m} w_k d_k \varphi_{lk}\right)^2 - \sum_{k\in s_m} w_k^2 d_k \varphi_{lk}^2\right)\hat{\sigma}_l^2. \qquad (3.6)$$

This difference can be relatively small, especially in business surveys characterized with asymmetric data. This is the case when $\max_{k\in s_m}\left(w_k d_k \varphi_{lk}\right) \cong \sum_{k\in s_m} w_k d_k \varphi_{lk}$. This is in line with the results shown by Mills et al. (2013).

Overall, the total variance can be considered as approximately linear in terms of the unit-level errors, especially in the case of sample surveys where $\hat{V}_{\text{ORD}}\left(\hat{t}_d\right), \hat{V}_{\text{DIF}}\left(\hat{t}_d\right),$ and $\hat{V}_{\text{MIX}}\left(\hat{t}_d\right)$ are significant contributors to the total variance.

4 Simulation study

The sum of the item contributions is expected to be close enough to the estimated variance due to non-response. Simulations were conducted to assess the validity of the proposed score. The goal was then to evaluate if the proposed item contribution is a good approximation of the real contribution to the total variance of a given unit. In order to do so, the total contributions of a random subset of s_m were compared to the difference of the estimated variances where this subset is respectively considered as nonresponding units and responding units.

The following steps explain how simulations were performed.

1. A population was created, starting from an auxiliary variable x generated according to a gamma distribution with a mean of 48 and a variance of 768. The variable of interest y was created conditionally on x from a gamma distribution with a mean of $1.5x$ and a variance of $16x.$ These parameters are the same as the ones set by Beaumont and Bissonnette (2011).
2. A simple random sample s was selected from this population and an independent non-response subset s_m was generated using Bernoulli sampling.
   a. The nonresponding units from s_m were imputed using ratio imputation, where $y_k^* = x_k\left(\sum_{l \in s_r} y_l\right)\left(\sum_{l \in s_r} x_l\right)^{-1}$ and
   $$\hat{\sigma}_k^2 = x_k \frac{\sum_{l \in s_r} (y_l - y_l^*)^2}{\sum_{l \in s_r} x_l}.$$
   b. The population total $\hat{t},$ the variance components $\hat{V}_\bullet\left(\hat{t}\right),$ and the unit-level decompositions $\delta_k\left(\hat{V}_\bullet\left(\hat{t}\right)\right)$ were estimated, where the subscript $\bullet$ represents any of the variance components.
3. A subset, $\Lambda,$ of units, $\lambda,$ from $s_m,$ independently selected from a Bernoulli experiment, was moved from s_m to s_r to simulate non-response conversion. Therefore, we have a new partition, $P_s^{(\Lambda)},$ with $s_m^{(\Lambda)} = s_m \setminus \Lambda$ and $s_r^{(\Lambda)} = s_r \cup \Lambda.$
   a. The nonresponding units k from $s_m^{(\Lambda)}$ were re-imputed using a ratio model which is given by $y_k^{**} = x_k\left(\sum_{l \in s_r^{(\Lambda)}} y_l\right)\left(\sum_{l \in s_r^{(\Lambda)}} x_l\right)^{-1}$ and
   $$\hat{\sigma}_k^{*2} = x_k \frac{\sum_{l \in s_r^{(\Lambda)}} (y_l - y_l^{**})^2}{\sum_{l \in s_r^{(\Lambda)}} x_l}.$$
   b. The population total, $\hat{t}^{(\Lambda)},$ and the variance components, $\hat{V}_\bullet\left(\hat{t}^{(\Lambda)}\right),$ were estimated.

4. The total of the unit-level decompositions, $\sum_{\lambda \in \Lambda} \delta_\lambda \left(\hat{V}_\bullet \left(\hat{t}\right)\right)$, for units λ from Λ was compared to the difference in the variance component estimates, $\hat{V}_\bullet \left(\hat{t}\right) - \hat{V}_\bullet \left(\hat{t}^{(\Lambda)}\right)$. The relative difference in the decomposition error, DRel, was calculated as

$$\text{DRel} = \frac{\left(\hat{V}_\bullet \left(\hat{t}^{(\Lambda)}\right) + \sum_{\lambda \in \Lambda} \delta_\lambda \left(\hat{V}_\bullet \left(\hat{t}\right)\right)\right) - \hat{V}_\bullet \left(\hat{t}\right)}{\hat{V}_\bullet \left(\hat{t}\right)}. \quad (4.1)$$

Steps 1 to 4 were independently repeated with different combinations of population size, sample size, response rate, and conversion rate as described in 4.1, 4.2 and 4.3.

4.1 Simulation scenario 1: Fixed parameters

In scenario 1, population size, sample size, response rate, and conversion rate were respectively set to 400, 100, 70%, and 33.3%, with 200 independent iterations. The results are shown in Figures 4.1 and 4.2.

Both Figures 4.1 and 4.2 show that the sum of the unit-level decomposition is a good predictor of the change in the non-response component estimates. The average relative difference in the variance estimates is low at 2.1%, but the standard error is large at 5.8%. Out of 200 relative differences, only 19 are not within the +/– 10% range but they are all above 10%. If a nonrespondent is converted to a respondent, we conclude that the non-response component of the variance will approximately be reduced by the measured contribution of this unit.

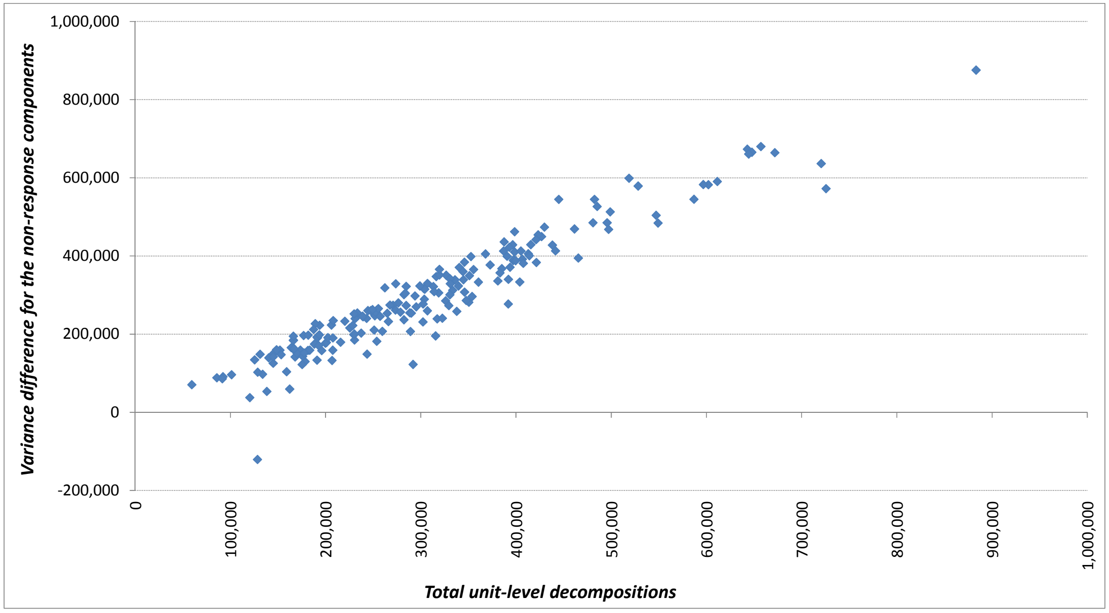


Figure 4.1 Variance difference for the non-response components versus total unit-level decompositions with fixed parameters.

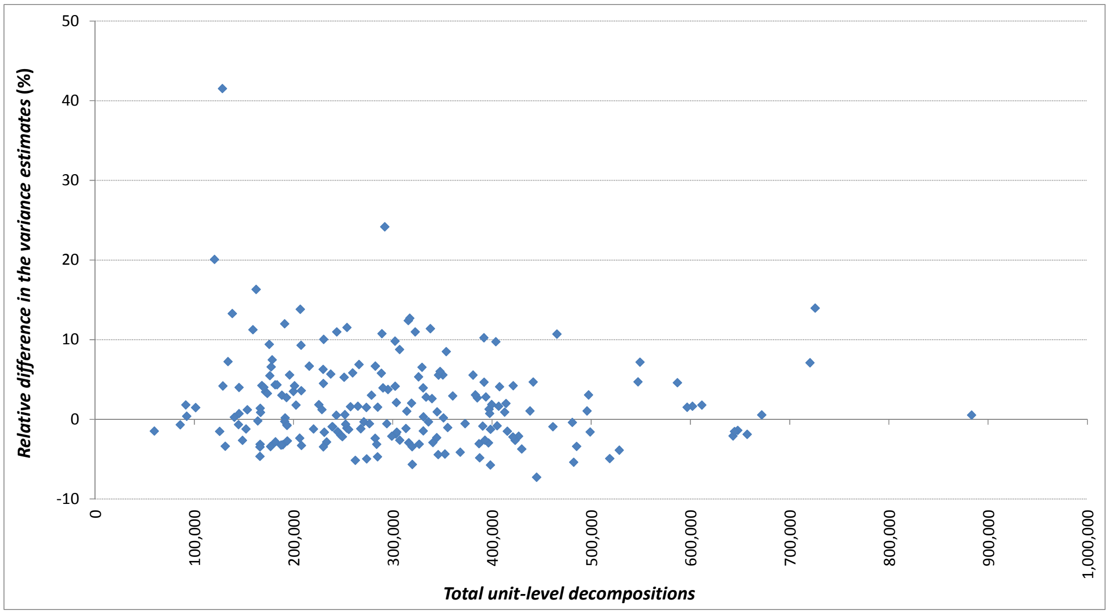


Figure 4.2 Relative difference in the variance estimates versus total unit-level decompositions with fixed parameters.

4.2 Simulation scenario 2: Varying population and sample sizes

In scenario 2, the population size ranged from 100 to 50,000, with sample rate, response rate, and conversion set to 20%, 70%, and 33.3% respectively. More iterations (40) were created for the smallest population $(N = 100)$, and less (10) for the largest $(N = 50{,}000)$, for operational considerations. The results are shown in Figures 4.3 and 4.4.

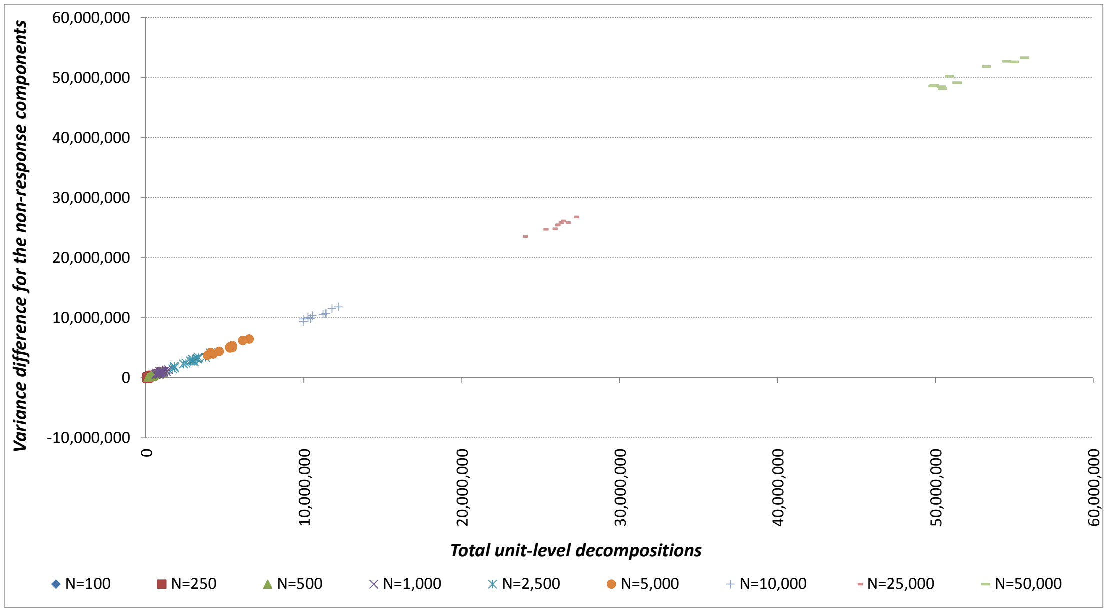


Figure 4.3 Variance difference for the non-response components versus total unit-level decompositions, varying population sizes.

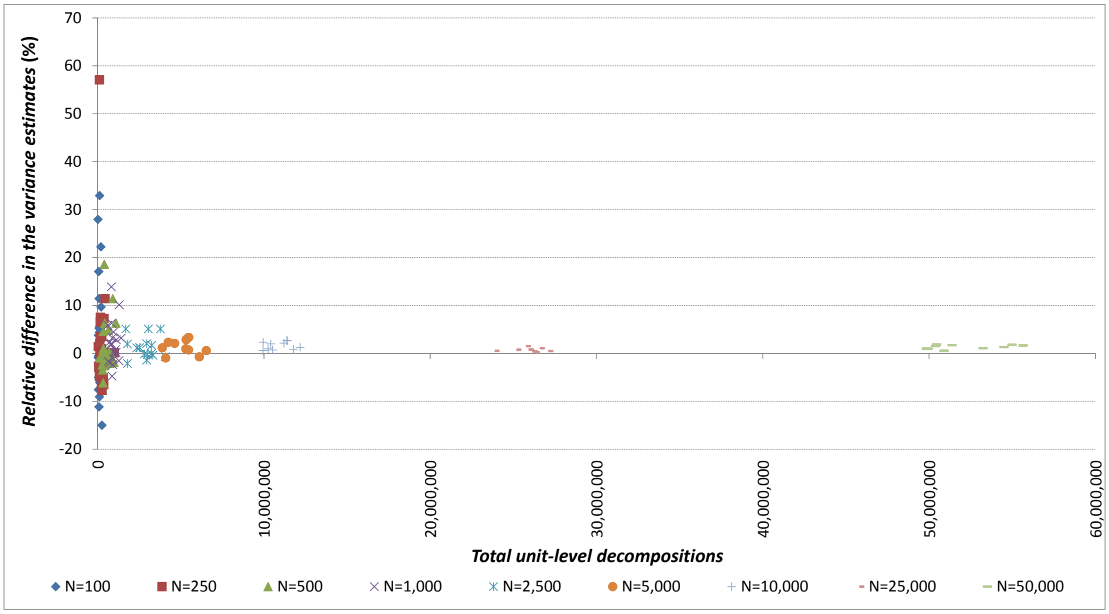


Figure 4.4 Relative difference in the variance estimates versus total unit-level decompositions, varying population sizes.

Both Figures 4.3 and 4.4 show that the relative differences in the decomposition errors are more volatile for smaller populations but rapidly converge close to 0 as population and sample sizes increase. This is further confirmed by Table 4.1.

Table 4.1
Count, average and standard deviation of relative differences in the variance estimates by population sizes

Population Size (N)	Relative Differences in the Variance Estimates in percentage		
	Count	Average	Standard Deviation
100	33(*)	2.2	10.6
250	30	1.6	11.4
500	25	1.0	5.3
1,000	20	2.2	4.4
2,500	10	1.2	2.3
5,000	10	1.2	1.4
10,000	10	1.6	0.8
25,000	10	0.7	0.4
50,000	10	1.3	0.4
Grand Total	163	1.6	7.3

(*): Out of 40 replicates created, only 33 had converted units.

To identify the sources of this instability, the relative differences between the estimated imputation variance, $\mathrm{DRel}\left(\hat{\sigma}_k^2\right) = \left(\hat{\sigma}_k^2 - \hat{\sigma}_k^{2(\Lambda)}\right)\big/\hat{\sigma}_k^{2(\Lambda)}$, and the relative difference between the estimated imputation relationship element, $\mathrm{DRel}\left(\varphi_{lk}\right) = \left(\varphi_{lk} - \varphi_{lk}^{(\lambda)}\right)\big/\varphi_{lk}^{(\lambda)}$, were measured for all units, $k \notin \Lambda$ and $l \notin \Lambda$. Note that under the ratio imputation model, both are constant for a given replicate, i.e., $\mathrm{DRel}\left(\hat{\sigma}_k^2\right) = \mathrm{DRel}\left(\hat{\sigma}^2\right)$ and $\mathrm{DRel}\left(\varphi_{lk}\right) = \mathrm{DRel}\left(\varphi\right)$. After the deletion of 2 extreme replicates, the correlation between the relative difference in the variance estimates DRel and $\mathrm{DRel}\left(\hat{\sigma}^2\right)$ is 0.78 while the correlation between DRel and $\mathrm{DRel}\left(\varphi\right)$ is 0.01. This illustrates that the instability is primarily caused by the variability of the $\hat{\sigma}_l^{(\lambda)} = \hat{\sigma}_l$ estimates. From this scenario, the conclusions are:

- Assumption 2 becomes valid for large enough sample sizes and leads to more accurate unit-level decomposition for consistent imputation model variance estimators.
- The unit-level decomposition is robust to assumption 3 validity.

4.3 Simulation scenario 3: Varying conversion rates

In scenario 3, the population and sample sizes were fixed to 2,500 and 500 respectively, and response rate is set to 50%. The conversion rates (CR) varied from 10% to 100% by increments of 10%, in order to generate different sizes of subset Λ, with 15 iterations each. The results are shown in Figures 4.5 and 4.6.

Both Figures 4.5 and 4.6 show that the relative difference in the decomposition errors becomes biased as the size of Λ increases, as confirmed in Table 4.2. This is primarily due to non-linearity of $\hat{V}_{\mathrm{NR}}\left(\hat{t}_d\right)$, as demonstrated in equation (3.6). The monotone nature of the relationship in Figure 4.5 suggests that the ordering of the error contributors is not affected, i.e., the large estimated contributors will have larger effect on the variance than the ones with a small estimated contribution.

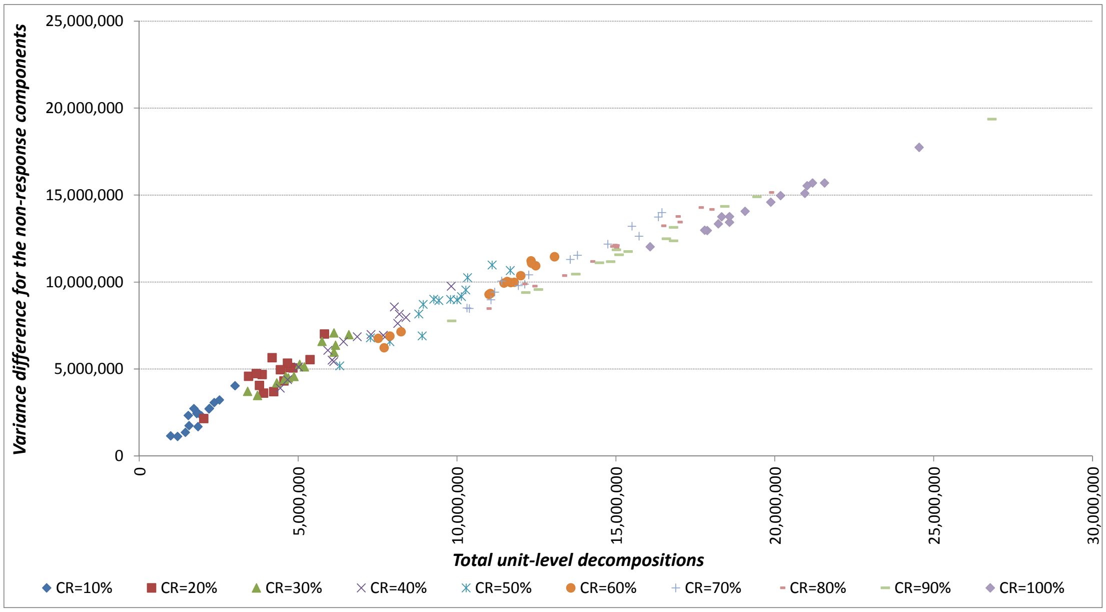


Figure 4.5 Variance difference for the non-response components versus total unit-level decompositions, varying conversion rates (CR).

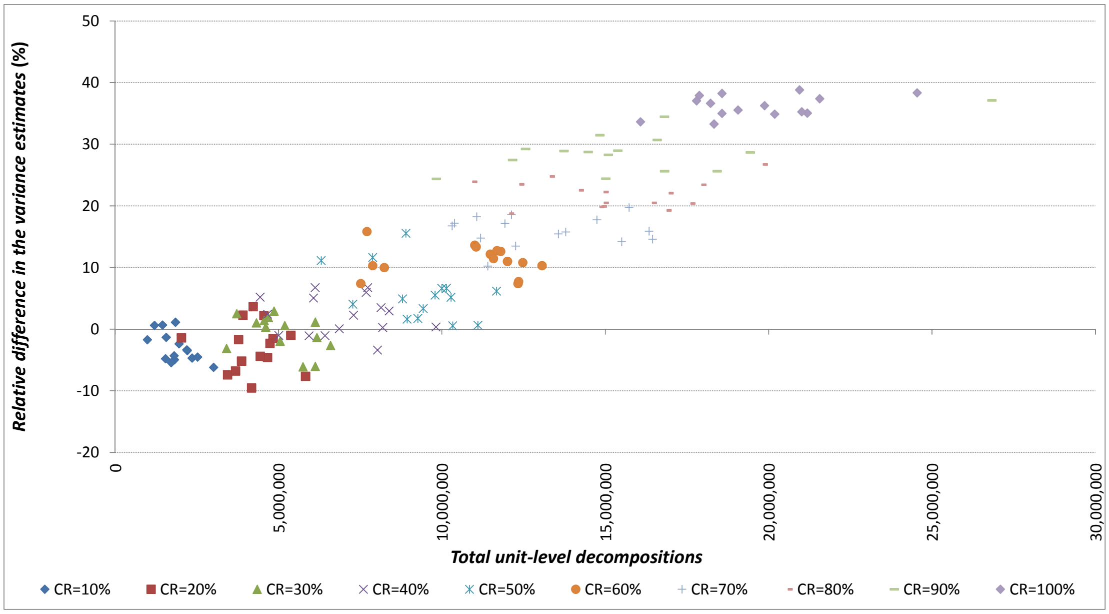


Figure 4.6 Relative difference in the variance estimates versus total unit-level decompositions, varying conversion rates (CR).

Table 4.2
Count, average, and standard deviation of relative differences in the variance estimates by conversion rates (CR)

Conversion Rate (CR)	Relative Differences in the Variance Estimates in percentage		
	Count	Average	Standard Deviation
10%	15	-3.0	2.4
20%	15	-3.0	3.9
30%	15	-0.5	3.0
40%	15	2.2	3.3
50%	15	5.7	4.3
60%	15	11.1	2.4
70%	15	16.0	2.4
80%	15	21.9	2.3
90%	15	28.9	3.5
100%	15	36.2	1.7
Grand Total	150	11.5	13.5

Despite the fact that the relative differences in the variance estimates are not null on average, it doesn't prevent the use of the proposed decomposition of errors to identify the largest sources of variance, especially in asymmetric populations. Mills et al. (2013) showed through a simulation how this could be successfully adapted into an efficient active collection strategy.

5 Conclusion

The proposed unit-level score is a good approximation of the unit impact on the variance due to non-response. It is applicable for different survey designs, compliant with calibration estimators for domain totals and works with many common imputation methods. The assumptions on which the decomposition relies are generally valid in common surveys using unbiased imputation methods and consistent estimators of imputation model parameters. The simulation results show that this approach becomes more accurate with larger sample sizes. The decomposition of the non-response variance is biased due to its non-linearity. However, the bias is smaller in asymmetric populations and when focusing on a small number of nonresponding units. The fact that the ordering of units using the estimated contribution to variance due to non-response is similar to the real order is an important aspect when the priority is to identify the largest contributors, not necessarily their actual contributions, to the total error.

This paper presented the method in a univariate context but it can be easily extended to a multivariate framework, using a distance function to combine the item contributions into a unit contribution. The idea remains to focus our attention in terms of collection treatments or manual verification on cases where the unit scores are the highest. In this case the non-response follow-up treatment might be different for unit non-response compared to partial non-response. For example, a telephone follow-up could be used to collect all the items for the total nonresponding units with the larger score; and the partial nonrespondents with a large score could be sent to an analyst for review, depending on the budget for follow-up. Moreover, if this score can be computed several times during the collection period, then non-response follow-ups will be more efficient because the unit score will be more accurate and the quality might become satisfactory for some estimates. Simulation results show that the proposed score is a good approximation to the contribution of a unit to the variance due to non-response. Subsequently, this score could be used to determine how many and which nonresponding units should be followed in order to reach a given estimated coefficient of variation.

This work was initially done for non-response prioritization under the Rolling Estimate iterative adaptive design process for IBSP. Following the original plan, key item estimates would be computed with their associated quality indicators at several specific times during the collection period. After each specific time, the units with the largest contributions according to our method would be prioritized for follow-up.

Acknowledgements

The authors want to thank the reviewers (Cynthia Bocci and Jessica Andrews), the associate editor, the referees and the assistant editor for their valuable feedback.

Appendix

Proof 1

$$\sum_{k \in s_m} \delta_k \left(\hat{V}_{\text{DIF}} \left(\hat{t}_d \right) \right) = \sum_{k \in s_m} (1 - \pi_k)\, w_k^2 d_k \hat{\sigma}_k^2 = \hat{V}_{\text{DIF}} \left(\hat{t}_d \right).$$

Proof 2

$$\begin{aligned}
\sum_{k\in s_m} \delta_k \left(\hat{V}_{\text{MIX}}\left(\hat{t}_d\right)\right) &= \sum_{k\in s_m}\left(2\sum_{l\in s_r} w_k d_k \varphi_{lk}\left(w_l - 1\right) d_l \hat{\sigma}_l^2 - 2w_k\left(w_k - 1\right) d_k \hat{\sigma}_k^2\right)\\
&= 2\sum_{k\in s_m}\sum_{l\in s_r} w_k d_k \varphi_{lk}\left(w_l - 1\right) d_l \hat{\sigma}_l^2 - 2\sum_{k\in s_m} w_k\left(w_k - 1\right) d_k \hat{\sigma}_k^2\\
&= 2\sum_{l\in s_r}\sum_{k\in s_m} w_k d_k \varphi_{lk}\left(w_l - 1\right) d_l \hat{\sigma}_l^2 - 2\sum_{k\in s_m} w_k\left(w_k - 1\right) d_k \hat{\sigma}_k^2\\
&= 2\sum_{l\in s_r}\left(\sum_{k\in s_m} w_k d_k \varphi_{lk}\right)\left(w_l - 1\right) d_l \hat{\sigma}_l^2 - 2\sum_{k\in s_m} w_k\left(w_k - 1\right) d_k \hat{\sigma}_k^2\\
&= 2\sum_{l\in s_r} W_{dl}\left(w_l - 1\right) d_l \hat{\sigma}_l^2 - 2\sum_{k\in s_m} w_k\left(w_k - 1\right) d_k \hat{\sigma}_k^2\\
&= \hat{V}_{\text{MIX}}\left(\hat{t}_d\right).
\end{aligned}$$

Proof 3

$$\begin{aligned}
\sum_{k\in s_m} \delta_k \left(\hat{V}_{\text{NR}}\left(\hat{t}_d\right)\right) &= \sum_{k\in s_m}\left(\sum_{l\in s_r}\left(2W_{dl} w_k d_k \varphi_{lk} - w_k^2 d_k \varphi_{lk}^2\right)\hat{\sigma}_l^2 + w_k^2 d_k \hat{\sigma}_k^2\right)\\
&= \sum_{k\in s_m}\left(\sum_{l\in s_r}\left(2W_{dl} w_k d_k \varphi_{lk} - w_k^2 d_k \varphi_{lk}^2\right)\hat{\sigma}_l^2\right) + \sum_{k\in s_m} w_k^2 d_k \hat{\sigma}_k^2\\
&= \sum_{l\in s_r}\left(\sum_{k\in s_m}\left(2W_{dl} w_k d_k \varphi_{lk} - w_k^2 d_k \varphi_{lk}^2\right)\hat{\sigma}_l^2\right) + \sum_{k\in s_m} w_k^2 d_k \hat{\sigma}_k^2\\
&= \sum_{l\in s_r}\left(2W_{dl}\sum_{k\in s_m} w_k d_k \varphi_{lk}\hat{\sigma}_l^2 - \sum_{k\in s_m} w_k^2 d_k \varphi_{lk}^2 \hat{\sigma}_l^2\right) + \sum_{k\in s_m} w_k^2 d_k \hat{\sigma}_k^2\\
&= \sum_{l\in s_r}\left(2W_{dl}^2\hat{\sigma}_l^2 - \sum_{k\in s_m} w_k^2 d_k \varphi_{lk}^2 \hat{\sigma}_l^2\right) + \sum_{k\in s_m} w_k^2 d_k \hat{\sigma}_k^2\\
&= \sum_{l\in s_r} 2W_{dl}^2\hat{\sigma}_l^2 - \sum_{l\in s_r}\sum_{k\in s_m} w_k^2 d_k \varphi_{lk}^2 \hat{\sigma}_l^2 + \sum_{k\in s_m} w_k^2 d_k \hat{\sigma}_k^2\\
&= \hat{V}_{\text{NR}}\left(\hat{t}_d\right) + \sum_{l\in s_r} W_{dl}^2\hat{\sigma}_l^2 - \sum_{l\in s_r}\sum_{k\in s_m} w_k^2 d_k \varphi_{lk}^2 \hat{\sigma}_l^2\\
&= \hat{V}_{\text{NR}}\left(\hat{t}_d\right) + \sum_{l\in s_r}\left(W_{dl}^2 - \sum_{k\in s_m} w_k^2 d_k \varphi_{lk}^2\right)\hat{\sigma}_l^2\\
&= \hat{V}_{\text{NR}}\left(\hat{t}_d\right) + \sum_{l\in s_r}\left(\left(\sum_{k\in s_m} w_k d_k \varphi_{lk}\right)^2 - \sum_{k\in s_m} w_k^2 d_k \varphi_{lk}^2\right)\hat{\sigma}_l^2.
\end{aligned}$$

References

Beaumont, J.-F., and Bissonnette, J. (2011). Variance estimation under composite imputation: The methodology behind SEVANI. *Survey Methodology*, 37, 2, 171-179. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2011002/article/11605-eng.pdf.

Beaumont, J.-F., and Bocci, C. (2009). Variance estimation when donor imputation is used to fill in missing values. *Canadian Journal of Statistics*, 37, 400-416.

Beaumont, J.-F., Bocci, C. and Haziza, D. (2014). An adaptive data collection procedure for call prioritization. *Journal of Official Statistics*, 30, 607-621.

Beaumont, J.-F., Haziza, D. and Bocci, C. (2011). On variance estimation under auxiliary value imputation in sample surveys. *Statistica Sinica*, 21, 515-537.

Biemer, P.P. (2010). Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly*, 74, 5, 817-848.

Bosa, K., and Godbout, S. (2014). *IBSP Quality Measures – Methodology Guide*. Business Survey Methods Division. Internal document.

Godbout, S., Beaucage, Y. and Turmelle, C. (2011). Achieving quality and efficiency using a top-down approach in the Canadian integrated business statistics Program. *Proceedings of the Conference of European Statisticians*. United Nations Statistical Commission and Economic Commission for Europe. Work Session on Statistical Data Editing. Ljubljana, Slovenia, 9-11 May 2011.

Groves, R.M., and Heeringa, S.G. (2006). Responsive design for household surveys: Tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society, Series A*, 169, No. 3, 439-457.

Mills, F., Godbout, S., Bosa, K. and Turmelle, C. (2013). Multivariate selective editing in the integrated business statistics program. *Proceedings of the Joint Statistical Meeting 2013 - Survey Research Methods Section*. August 2013. Montréal, Canada.

Särndal, C.-E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, 18, 2, 241-252. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/1992002/article/14483-eng.pdf.

Schouten, B., Calinescu, M. and Luiten, A. (2013). Optimizing quality of response through adaptive survey designs. *Survey Methodology*, 39, 1, 29-58. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2013001/article/11824-eng.pdf.

Statistics Canada (2015). *Integrated Business Statistics Program Overview*. Statistics Canada Catalogue no. 68-515-X. Ottawa.

Turmelle, C., Godbout, S. and Bosa, K. (2012). Methodological challenges in the development of Statistics Canada's new integrated business statistics program. *Proceedings of the International Conference on Establishment Surveys IV*. Montréal, Canada.

Vol. 44, No. 2, pp. 309-338
Statistics Canada, Catalogue No. 12-001-X

Comparison of the conditional bias and Kokic and Bell methods for Poisson and stratified sampling

Thomas Deroyon and Cyril Favre-Martinoz[1]

Abstract

In business surveys, it is common to collect economic variables with highly skewed distribution. In this context, winsorization is frequently used to address the problem of influential values. In stratified simple random sampling, there are two methods for selecting the thresholds involved in winsorization. This article comprises two parts. The first reviews the notations and the concept of a winsorization estimator. The second part details the two methods and extends them to the case of Poisson sampling, and then compares them on simulated data sets and on the labour cost and structure of earnings survey carried out by INSEE.

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

1 Introduction

In survey statistics, a population unit is influential if the estimators produced on a sample drawn from that population change significantly depending on whether or not that unit is sampled. The concept of an influential unit therefore depends on several factors, which determine what Beaumont, Haziza and Ruiz-Gazen (2013) called a configuration:

- a sampling design for a population;
- one or more variables of interest and a parameter of interest on the distribution of this variable;
- an estimator calculated on the sample for this parameter of interest.

A unit may be influential in one configuration and not in another. For example, it can have a significant effect on the estimator of the total of a variable in a particular domain, but have only a minor influence on the estimator of the total of that variable in the total population.

Chambers (1986) distinguishes two types of influential units: non-representative atypical values are units that have provided erroneous information or are found in these exceptional situations. The information collected on these units cannot be extrapolated to the rest of the population; these units are usually identified during collection or during control of the data collected and processed *via* specific procedures (for answers considered to be erroneous, the information collected is, for example, replaced by a missing and imputed value. It can also be corrected by recontacting the unit in question. For units that are in an exceptional situation and for which we are sure the case is unique, it is common to put their weight to 1).

Representative influential units provided correct answers and are not a priori unique in the population. They are common in surveys of businesses, a population for which many variables have a very skewed distribution. In particular, the variables reflecting volumes or amounts (turnover, value added, payroll,

[1] Thomas Deroyon, INSEE, Paris, France; Cyril Favre-Martinoz, INSEE, Saint-Denis de la Réunion, France. E-mail: cyril.favre-martinoz@insee.fr.

investment, energy consumption, research and development expenditure and anti-pollution expenditure to name some of the key variables of INSEE business surveys) are characterized by a high concentration of low values, corresponding to many small businesses, and some very high values associated with large or very large businesses.

To limit the effects of this wide dispersion of the variables of interest in the population of businesses, the classical sampling design applied to them is a stratified design, in which size, measured in number of employees, is used as a stratification variable. In most cases, this makes it possible to assign businesses inclusion probabilities correlated with the amounts they reported in the survey. In these designs, large businesses are surveyed exhaustively, as are businesses that, according to the auxiliary information available in the sampling frames, are likely to report very large amounts in the survey, regardless of their size.

In practice, however, it is impossible to be entirely protected against influential observations at the sampling design stage. Indeed, the information in the sampling frames may be affected by measurement errors. For example, the number of employees in the sampling frames is a variable derived from returns to social security organizations that requires a significant amount of controls and adjustments and takes two years to reach a definitive value for a given year. It is thus possible, when drawing a sample, to use the last known definitive value, but which relates to a business's previous situation, or to use the nearest preliminary value, which will be affected by more measurement errors. In both cases, the variable used for stratification may not correspond to the actual situation of the business at the time of the survey, creating businesses sampled in the wrong stratum (called "strata jumpers"), whose sampling weight is much too high compared with their survey responses.

The auxiliary variables available to define the sampling designs may also be only weakly correlated with the survey themes. It is therefore difficult to identify businesses that are innovative or involved in research and development activities based solely on their industry, size, region of establishment, duration of existence or legal category. The same goes for the amounts invested in sustainable development (measured in France by the Antipol Survey conducted by INSEE: https://www.insee.fr/fr/metadonnees/source/s1232).

Surveys can also collect several weakly correlated variables of interest. The sampling design, which aims to achieve the highest possible precision for the survey's core variables of interest, may not be appropriate for other, less significant variables, e.g., the portion of turnover generated by online sales. In particular, some businesses that report atypical values for secondary variables of interest in the survey may not have been identified and placed in a comprehensive stratum.

Finally, many business surveys are conducted at regular intervals, most often every year, and aim to estimate both the annual levels of the main variables of interest and their evolution. To meet these two objectives, the sample surveyed in the non-exhaustive strata is not renewed in full each year, but a portion is retained. For example, the sample of business surveys on Information and Communication Technologies (ICT-E) is renewed by half each year; businesses sampled in a given year are surveyed two years in a row (see Demoly, Fizzala and Gros, 2014). In this case, the businesses retain the sampling weight with which they were initially sampled, which may no longer match their characteristics at the time of the survey, resulting in the appearance of “stratum jumpers” and potentially influential units.

Classical estimators in the presence of survey data (for example, the expansion estimator or the estimator adjusted for total non-response) have (virtually) no bias but can be very unstable in the presence of influential values. Robust estimation methods must then be implemented to limit their impact. The principle of these methods is to modify the estimation weights or the declared values by the influential units in order to make the estimators more stable, at the risk of biasing them. More precisely, the estimators to which these methods lead must have a mean square error significantly lower than that of classical expansion estimators in the presence of influential data, without losing too much efficiency in the absence of atypical values in the sample. The processing of influential values therefore lies in a compromise between bias and variance.

The most common method in practice for dealing with the problem of influential values is winsorization, which applies to estimating totals of variables of interest. For a given variable of interest, this consists of partitioning the sample and associating each part of the sample with a threshold; for example, in the case of a sample selected by stratified simple random sampling, the sample is cut according to the drawing strata, and a different threshold is associated with each stratum. Units in the sample for which the values of the variable are greater than the threshold associated with their part of the sample have their response or their weight decreased, while the responses and weights of the other units are not modified. There are two forms of winsorization in the literature, which differ depending on how the variable or weight is modified when the variable of interest exceeds the threshold. In standard winsorization, also known as Type I winsorization, values that exceed the threshold are truncated at the threshold. In this article, we will use the form proposed by Dalén (1987) and Tambay (1988), also called Type II winsorization, because it ensures that winsorized weights greater than 1 are obtained. This method will be briefly reviewed in Section 2.

In the application of winsorization, the choice of thresholds is crucial; a bad choice can lead to winsorized estimators with a higher mean square error than the classical estimators via the introduction of a very high bias that is difficult to correct later. The choice of these thresholds has been the subject of numerous studies, including by Kokic and Bell (1994), Rivest and Hurtubise (1995) and Favre-Martinoz, Haziza and Beaumont (2015). In the case of a simple random stratified design without replacement, Kokic and Bell (1994) determined the theoretical formulas and algorithms for calculating the thresholds that realizations in the winsorized estimator with the lowest mean square error possible, under the hypothesis that the realizations of the variable of interest are identically distributed in each stratum, the mean square error being calculated under the sampling design and the law of the variable of interest. In the case of repeated surveys, they suggest using historical data collected in previous editions of the surveys to calculate these thresholds. Clark (1995) generalized the results of Kokic and Bell (1994) in the case of a ratio estimator by calculating the mean square error with respect to the model only.

Other methods have been proposed for identifying and processing influential units in survey statistics. One of these, introduced by Beaumont et al. (2013), is based on the concept of conditional bias, a measure of influence proposed by Moreno-Rebollo, Muñoz-Reyez and Muñoz-Pichardo (1999) and

Moreno-Rebollo, Muñoz-Reyez, Jimenez-Gamero and Muñoz-Pichardo (2002). Unlike the winsorization methods mentioned above, which are only suitable for certain sampling designs and require fairly rich information outside the sample, the method proposed by Beaumont et al. (2013) can be applied a priori to any sampling design and uses only the survey responses. However, it does not necessarily lead to the processed estimator of influential units with the smallest mean square error, but to the estimator on which the influence of the most influential unit is the lowest in absolute value. Favre-Martinoz et al. (2015) and Favre-Martinoz, Haziza and Beaumont (2016) proposed adaptations of the conditional bias method for calculating winsorization thresholds and factoring in an additional sampling phase and estimation in domains.

The purpose of this paper is to compare the efficiency of the winsorization and conditional bias methods to treat influential values. In Section 2, we review the winsorization method and the calculation of winsorization thresholds proposed by Kokic and Bell in stratified simple random sampling. We also propose an extension of the Kokic and Bell method for a Poisson sampling design. After briefly reviewing the principles of robust estimation based on conditional bias in Section 3, we present in Section 4 simulations to compare the extension of the Kokic and Bell method with the conditional bias methods in the Poisson case. Finally, an example of the practical application of the Kokic and Bell method and its extension to the Poisson case is presented in Section 4, which compares them with a method based on conditional biases in the context of the labour cost and structure of earnings survey carried out by INSEE.

2 The processing of influential units by winsorization following the approach of Kokic and Bell

In this section, we present the method initially proposed by Kokic and Bell (1994), which applies to samples selected through stratified simple random sampling, and an extension of this method to the case of samples selected through Poisson sampling.

2.1 Case of stratified simple random sampling

Consider a finite is a population U of size N and a variable of interest X observed on a sample S of fixed size n and for which we are looking to estimate the total $T(X) = \sum_{i \in U} X_i$ on the population. The approach of Kokic and Bell (1994) is based on the following hypotheses:

- X is a positive or nil variable;
- S is selected according to a stratified simple random sampling design $P,$ following strata $U_h,\ h = 1, \ldots, H.$ In each stratum of size $N_h,$ a sample S_h of size n_h is selected according to a simple random design without replacement. The expectation with respect to the sampling design will be denoted E_p afterwards;
- in each stratum $U_h,$ the values of X in the population are derived from random variables X_{hi} that are independent and identically distributed according to a law $\mathcal{L}_h$ (or of the same model m)

with expectation μ_h. The expectation and the variance with respect to this model will be denoted E_m and V_m respectively hereafter;

- we have, for each stratum U_h, N_h realizations $\breve{X}_{hi}$ of the variable X derived from the same law $\mathcal{L}_h$ but independent of the sample S_h.

In this context, Kokic and Bell (1994) propose applying a Type II winsorization; they associate with each stratum U_h a threshold K_h independent of the sample S and define the winsorized variable $\tilde{X}$, for $i \in S$, by:

$$\tilde{X}_{hi} = \begin{cases} X_{hi} & \text{if } X_{hi} < K_h \\ \dfrac{n_h}{N_h} X_{hi} + \left(1 - \dfrac{n_h}{N_h}\right) K_h & \text{if } X_{hi} \geq K_h. \end{cases}$$

The winsorized estimator of the total X is then the expansion estimator of the total of the winsorized variable $\tilde{X}$: $\hat{T}\left(\tilde{X}\right) = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i \in S_h} \tilde{X}_{hi}$.

The thresholds K_h are determined so as to obtain the estimator $\hat{T}\left(\tilde{X}\right)$ with the lowest mean square error with respect to both the sampling design and the law of X in each stratum, i.e.,

$$(K_h^*)_{h=1,\ldots,H} \in \text{Argmin}_{(K_h)_{h=1,\ldots,H}} E_m E_P \left\{\left[\hat{T}\left(\tilde{X}\right) - T\left(X\right)\right]^2\right\}.$$

The optimal thresholds must therefore protect the winsorized estimator on average over all possible samples in the population, and on average on the law of the variable of interest, i.e., on average over all the possible populations considering the law of X.

Kokic and Bell (1994) place themselves in an asymptotic framework by considering a set of populations, sampling designs and samples indexed by $\nu \in \mathbb{N}$ such as:

- $\forall \nu \in \mathbb{N}, \forall h = 1, \ldots, H, n_{h_\nu} > 1;$
- $N_\nu, n_\nu \underset{\nu \to +\infty}{\longrightarrow} +\infty;$
- $\exists \epsilon \in \left]0, 1/2\right[, \forall \nu \in \mathbb{N}, \forall h = 1, \ldots, H, \epsilon < \frac{n_{h_\nu}}{N_{h_\nu}} < 1 - \epsilon;$
- the number of strata H is fixed.

They also propose denoting $J_{hi} = \mathbb{I}\left(X_{hi} \geq K_h\right)$ the winsorization indicator. To reduce the notations, we will omit in the rest of the article the indicator ν as well as the indicator i in the expression of the expectations and variances E_m and V_m of the random variables and X_{hi} J_{hi} under the law of X in the stratum h. Insofar as these variables are assumed to be independent and identically distributed in each stratum, $E_m\left(X_{hi}\right)$ for example, is indeed the same, regardless of the observation considered.

In this context, Kokic and Bell (1994) show that, at the optimum and asymptotically, all the thresholds are linked to one another by the relation:

$$\left(\frac{N_h}{n_h} - 1\right)\left(K_h - \mu_h\right) \sim -B \tag{2.1}$$

with $B = \sum_{h=1}^{H} N_h \left(1 - \frac{n_h}{N_h}\right)[K_h E_m (J_h) - E_m (X_h J_h)]$ the bias of the winsorized estimator. The notation $\sim$ corresponds to an asymptotic equivalence when n_ν tends toward infinity (which is equivalent to saying when ν tends toward infinity).

If we denote $X_{hi}^* = \left(\frac{N_h}{n_h} - 1\right)(X_{hi} - \mu_h)$ and $L = -B,$ then we can notice that at the optimum given (2.1), $J_{hi} = J_{hi}^* = \mathbb{I}(X_{hi}^* \geq L)$ and the bias B is the opposite of the zero-point of the function F defined by:

$$F(L) = L\left\{1 + \sum_{h=1}^{H} n_h E_m (J_h^*)\right\} - \sum_{h=1}^{H} n_h E_m (X_h^* J_h^*). \tag{2.2}$$

Determining the zero-point of the function F requires estimates of $\mu_h, E_m (J_h^*)$ and $E_m (X_h^* J_h^*).$ To do this, Kokic and Bell (1994) rely on observations of the variable X in each stratum. These observations must come from a source independent of the sample, since the demonstration of formulas (2.1) and (2.2) is based on the fact that the thresholds K_h are assumed to be independent of the sample $S.$

If we assume that for each stratum h we have p_h realizations $\breve{X}_{hi}$ of $X,$ then we can estimate F by:

$$\begin{aligned} \hat{F}(L) &= L\left\{1 + \sum_{h=1}^{H} n_h \frac{\sum_{i=1}^{p_h} \mathbb{I}\left(\breve{X}_{hi}^* \geq L\right)}{p_h}\right\} \\ &\quad - \sum_{h=1}^{H} n_h \frac{\sum_{i=1}^{p_h} \breve{X}_{hi}^* \mathbb{I}\left(\breve{X}_{hi}^* \geq L\right)}{p_h} \end{aligned} \tag{2.3}$$

with

$$\breve{X}_{hi}^* = \left(\frac{N_h}{n_h} - 1\right)\left(\breve{X}_{hi} - \frac{\sum_{j=1}^{p_h} \breve{X}_{hj}}{p_h}\right)$$

and estimate the optimal bias B as the opposite of the zero-point of $\hat{F}.$

Now, $\hat{F}$ is an increasing function and is linear by sections, which admits only one zero-point. This can be estimated simply by denoting $\breve{X}_{(i)}^*$ the values of $\breve{X}_{hi}^*$ sorted in ascending order and by calculating $\hat{F}\left(\breve{X}_{(1)}^*\right), \hat{F}\left(\breve{X}_{(2)}^*\right), \ldots$ until $\hat{F}$ sign changes.

Indeed, $\hat{F}\left(\breve{X}_{(1)}^*\right) = \breve{X}_{(1)}^* + \sum_{h=1}^{H} \frac{\sum_{i=1}^{p_h}\left(\breve{X}_{(1)}^* - \breve{X}_{hi}^*\right)}{p_h}$ is negative because $\breve{X}_{(1)}^*$ is by definition lower than all the others $\breve{X}_{hi}^*$ and because $\breve{X}_{(1)}^*$ is negative, since $\frac{\sum_{j=1}^{p_h} \breve{X}_{hj}^*}{p_h} = 0.$ However, $\hat{F}\left(\breve{X}_{(p)}^*\right) = \breve{X}_{(p)}^* \geq 0,$ for similar reasons by denoting $p = \sum_{h=1}^{H} p_h.$

By denoting j the indicator such as $\hat{F}\left(\breve{X}_{(j)}^*\right) \leq 0$ and $\hat{F}\left(\breve{X}_{(j+1)}^*\right) \geq 0,$ B can be estimated by linear interpolation, i.e., by

$$\hat{B} = -\frac{\breve{X}_{(j)}^* \hat{F}\left(\breve{X}_{(j)}^*\right) - \breve{X}_{(j+1)}^* \hat{F}\left(\breve{X}_{(j+1)}^*\right)}{\hat{F}\left(\breve{X}_{(j)}^*\right) - \hat{F}\left(\breve{X}_{(j+1)}^*\right)}. \tag{2.4}$$

2.2 Extension to the case of the Poisson sampling design

We now place ourselves in the situation in which the sampling design P by which S is selected is a Poisson sampling design, in which each unit i of the population can belong to the sample with a probability $\pi_i > 0.$ We are always interested in estimating the total in the population $T(X) = \sum_{i \in U} X_i$ of a variable $X.$ The extension of the Kokic and Bell method to this sampling design assumes:

- that X is a positive or nil variable;
- that it is possible to partition the population and the sample into subpopulations U_h and S_h in which all the values $d_{hi} X_{hi}$ are independent realizations from the same model verifying:

$$\forall h = 1, \ldots, H, \forall i \in U_h, d_{hi} \ X_{hi} = \mu_h + \epsilon_{hi}, \tag{2.5}$$

with

$$\begin{cases} E_m(\epsilon_{hi}) & = 0 \\ V_m(\epsilon_{hi}) & = \sigma_h^2 < +\infty \end{cases}$$

where E_m and V_m designates the expectation and variance with respect to the model (2.5).

In this context, we propose, as in the original method applied to stratified simple random sampling, associating a threshold $K_h, h = 1, \ldots, H$ with each part $S_h, h = 1, \ldots, H$ and defining:

- the winsorized variable $\tilde{X}$ by

$$\tilde{X}_{hi} = \begin{cases} X_{hi} & \text{if } \ d_{hi} X_{hi} \leq K_h \\ \dfrac{X_{hi}}{d_{hi}} + \left(1 - \dfrac{1}{d_{hi}}\right) \dfrac{K_h}{d_{hi}} & \text{if } \ d_{hi} X_{hi} > K_h, \end{cases} \tag{2.6}$$

where $d_{hi} = \frac{1}{\pi_i}$ is the weight of the unit i in part $h.$

- the winsorized estimator of the total X as the usual expansion estimator of the total $\tilde{X}$:

$$\hat{T}(\tilde{X}) = \sum_{h=1}^{H} \sum_{i \in S_h} d_{hi} \ \tilde{X}_{hi}. \tag{2.7}$$

In the article by Kokic and Bell (1994), the subpopulations with which the thresholds are associated are the drawing strata, which respect two properties: the draws are independent between strata, and the authors postulate an identical population model for all observations in the same stratum. In the case of Poisson sampling, the drawings are by nature independent between units.

The strong hypothesis underlying model (2.5) is that values X_{hi} multiplied by weights d_{hi} are assumed to have constant expectation in each stratum. This means that the inclusion probabilities within each stratum are defined proportionally to the variable of interest $X.$ In practice, these inclusion probabilities are often defined proportionally to a known auxiliary variable that is strongly correlated with X, which makes it possible to be close to the hypothesis underlying model (2.5).

Note also that model (2.5) is the one under which the Horvitz-Thompson estimator is optimal in the sense of minimizing the mean square error with respect to the model.

In the following, the random variables $d_{hi} X_{hi}$ being assumed to be independent and identically distributed within each stratum, we will denote $Z_{hi} = d_{hi} X_{hi}$.

We also place ourselves in the same asymptotic framework as Kokic and Bell (1994) by adapting the hypothesis on the inclusion probabilities:

$$\forall h = 1, \ldots, H, \exists\, (\lambda_{1h}, \lambda_{2h}) \in \,]0, 1[^2, \text{ such that } \forall i \in U_h, \min(\pi_i) > \lambda_{1h} \text{ and } \max(\pi_i) < \lambda_{2h}. \tag{2.8}$$

As in the approach presented in the previous section, the thresholds K_h are determined so as to minimize the mean square error of the winsorized estimator $\hat{T}(\tilde{X})$ with respect to both the model of the variable X and the sampling design P, i.e., on average across all possible populations, given the super-population model applied to X and on average for all samples drawn from these populations, given the sampling design P:

$$(K_h^*)_{h=1,\ldots,H} \in \operatorname{Argmin}_{(K_h)_{h=1,\ldots,H}} E_m E_P \left\{ \left[\hat{T}(\tilde{X}) - T(X) \right]^2 \right\}.$$

It is possible to show (see Appendix A) that at the optimum and asymptotically, denoting as previously $J_{hi} = \mathbb{I}(Z_{hi} > K_h)$ and omitting the indicator i in the expression of expectations and variances under model (2.5) of the variables Z_{hi} and J_{hi}:

$$\forall h = 1, \ldots, H, K_h \sim -\frac{A_h}{C_h + D_h} B \tag{2.9}$$

with

$$\begin{cases} A_h &= \sum_{i \in U_h} \frac{1}{d_{hi}} \left(1 - \frac{1}{d_{hi}}\right) \\ C_h &= \sum_{i \in U_h} \left(\frac{1}{d_{hi}}\right)^2 \left(1 - \frac{1}{d_{hi}}\right)^2 \\ D_h &= \sum_{i \in U_h} \frac{1}{d_{hi}} \left(1 - \frac{1}{d_{hi}}\right)^3 \end{cases}$$

and

$$B = \sum_{h=1}^{H} A_h \left[K_h E_m (J_h) - E_m (J_h Z_h) \right]. \tag{2.10}$$

B is the bias of the optimal winsorized estimator $\hat{T}(\tilde{X})$ at the optimum the threshold K_h is therefore equal to a near positive term, in contrast to the bias multiplied by the term $\frac{A_h}{C_h + D_h}$.

If we denote $L = -B$ and $X_{hi}^* = \frac{C_h + D_h}{A_h} Z_{hi}$, then asymptotically $J_{hi} = J_{hi}^* = \mathbb{I}(X_{hi}^* > L)$ using relation (2.9).

By injecting equivalence relation (2.9) into formula (2.10) defining $B,$ we obtain only optimally and asymptotically, B is the opposite of the zero-point of the function F defined by:

$$F(L) = L\left(1 + \sum_{h=1}^{H} \frac{A_h^2}{C_h + D_h} E_m(J_h^*)\right) - \sum_{h=1}^{H} \frac{A_h^2}{C_h + D_h} E_m(J_h^* X_h^*). \tag{2.11}$$

As in the previous section, we assume finally that we have, for each subpopulation $h,$ of p_h realizations $\breve{X}_{hi}$ drawn from the law of X and independent of the sample $S.$ With these observations, we can estimate F by:

$$\hat{F}(L) = L\left(1 + \sum_{h=1}^{H} \frac{A_h^2}{C_h + D_h} \frac{\sum_{i=1}^{p_h} \mathbb{I}\left(\breve{X}_{hi}^* > L\right)}{p_h}\right) - \sum_{h=1}^{H} \frac{A_h^2}{C_h + D_h} \frac{\sum_{i=1}^{p_h} \breve{X}_{hi}^* \mathbb{I}\left(\breve{X}_{hi}^* > L\right)}{p_h} \tag{2.12}$$

and estimate B by the opposite of the zero-point of $\hat{F}.$

We will denote $\breve{X}_{(j)}^*$ the values of the $\breve{X}_{hi}^*$ placed in ascending order. Then, between two successive values $\breve{X}_{(j)}^*$ and $\breve{X}_{(j+1)}^*,$ the indicators $\mathbb{I}\left(\breve{X}_{hi}^* > L\right),$ as functions of $L,$ remain constant and with a positive slope. $\hat{F}$ is therefore a linear and increasing function of $L.$

In addition, $\hat{F}(0) = -\sum_{h=1}^{H} \frac{A_h^2}{C_h+D_h} \frac{\sum_{i=1}^{p_h} \breve{X}_{hi}^*}{p_h} \le 0$ and, when L exceeds $\breve{X}_{(p)}^*,$ with $p = \sum_{h=1}^{H} p_h,$ $\hat{F}(L) = L \ge 0.$ To determine the zero-point of $\hat{F},$ it is necessary to operate using a method similar to that proposed by Kokic and Bell (1994) in the case of stratified simple random sampling:

- calculate $\hat{F}(0), \hat{F}\left(\breve{X}_{(1)}^*\right), \hat{F}\left(\breve{X}_{(2)}^*\right), \ldots, \hat{F}\left(\breve{X}_{(p)}^*\right);$
- identify the value j such as $\hat{F}\left(\breve{X}_{(j)}^*\right) \le 0$ and $\hat{F}\left(\breve{X}_{(j+1)}^*\right) \ge 0,$ assuming that $\breve{X}_{(0)}^* = 0;$
- B is then estimated by interpolation, as in the previous section:

$$\hat{B} = -\frac{\breve{X}_{(j)}^* \hat{F}\left(\breve{X}_{(j)}^*\right) - \breve{X}_{(j+1)}^* \hat{F}\left(\breve{X}_{(j+1)}^*\right)}{\hat{F}\left(\breve{X}_{(j)}^*\right) - \hat{F}\left(\breve{X}_{(j+1)}^*\right)}.$$

3 Review of methods based on conditional bias

3.1 Definition

The conditional bias of an estimator $\hat{\theta}$ for the parameter $\theta,$ for a unit $i \in U$ was defined in the framework of Sampling Theory by Moreno-Rebollo et al. (1999) as follows:

$$B_{1i}^{\hat{\theta}} = E_P\left(\hat{\theta} - \theta \,\middle|\, I_i = 1\right), \tag{3.1}$$

$$B_{0i}^{\hat{\theta}} = E_P\left(\hat{\theta} - \theta \,\middle|\, I_i = 0\right). \tag{3.2}$$

The conditional bias of a sampled unit is equal to the average of the difference between $\hat{\theta}$ and θ on the set of samples containing that unit. Similarly, the conditional bias of an unsampled unit is equal to the average of the sampling error for all samples not containing that unit.

In the case of a one-phase sampling design, the conditional bias of the Horvitz-Thompson estimator $\hat{T}(X) = \sum_{i \in S} \frac{x_i}{\pi_i}$ associated with a sampled unit i is defined by

$$B_{1i}^{\hat{T}(X)} = \sum_{j \in U} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) x_j \tag{3.3}$$

where π_{ij} designates the joint inclusion probability of units i and j in the sample. Conditional bias (3.3) is, in general, unknown since the values of the variable of interest are only observed for the units in the sample. In practice, it is possible to estimate it without bias, or in a robust way, from the sample. We consider the conditionally unbiased estimator (see, for example, Beaumont et al., 2013):

$$\hat{B}_{1i}^{\hat{T}(X)} = \sum_{j \in S} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_j \pi_{ij}} \right) x_j. \tag{3.4}$$

This estimator is conditionally unbiased in the sense that $E_P \left(\hat{B}_{1i}^{\hat{T}(X)} \middle| I_i = 1 \right) = B_{1i}^{\hat{T}(X)}$ only if π_{ij} are strictly positive. Moreover, conditional bias (3.3) and its estimator (3.4) depend on the inclusion probabilities π_i and the joint inclusion probabilities π_{ij}. In other words, conditional bias is a measure that takes the sampling design into account.

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

$$B_i^{\hat{T}(X)} (I_i = 1) = (d_i - 1) x_i. \tag{3.5}$$

Unlike the case of other sampling designs, such as simple random sampling without replacement, conditional bias (3.5) is known directly for all sample units and does not require estimation from the sample because it does not depend on any parameter of the finite population.

Conditional bias, as demonstrated by Beaumont et al. (2013), is a direct measure of the influence of each unit on the estimation error, the second relation being verified for maximum entropy sampling designs:

$$V\left[\hat{T}(X)\right] = \sum_{i \in U} B_{1i}^{\hat{T}(X)} y_i \tag{3.6}$$

$$\hat{T}(X) - T(X) \approx \sum_{i \in S} B_{1i}^{\hat{T}(X)} + \sum_{i \in U - S} B_{0i}^{\hat{T}(X)}. \tag{3.7}$$

3.2 A robust estimator based on conditional bias

As shown by formulas (3.6) and (3.7), the conditional bias (CB) measures the effect of each unit on the estimation error and the estimation variance. A robust estimator should be defined in such a way that observations of the sample have only controlled and limited values of their conditional bias. Based on this idea, Beaumont et al. (2013) suggested using an estimator of the form:

$$\begin{aligned}\hat{T}^{\text{CB}}(X)(c) &= \hat{T}(X) + \sum_{i\in S}\Psi_c\left[\hat{B}_{1i}^{\hat{T}(X)}\right] - \sum_{i\in S}\hat{B}_{1i}^{\hat{T}(X)}\\ &= \hat{T}(X) - \sum_{i\in S}\left[\hat{B}_{1i}^{\hat{T}(X)} - \Psi_c\left(\hat{B}_{1i}^{\hat{T}(X)}\right)\right]\end{aligned}$$

with Ψ_c the Huber function defined by

$$\Psi_c(t) = \begin{cases} c & \text{if } \; t \geq c \\ t & \text{if } \; -c < t < c \\ -c & \text{if } \; -c \leq t \end{cases}$$

and $\hat{B}_{1i}^{\hat{T}(X)}$ the estimator defined in (3.4).

The Huber function is used to limit the influence of the most influential units by truncating their conditional bias. Parameter c can be chosen according to various optimization criteria for the robust estimator. For example, c can be chosen to obtain the estimate having, under the sample design, the smallest mean square error. However, it is relatively complex or sometimes impossible to obtain an analytical expression of c for a given sample design.

Beaumont et al. (2013) suggest choosing $c^* \in \operatorname{argmin}_c \operatorname{argmax}_i \left|\hat{B}_{1i}^{\hat{T}^{\text{CB}}(X)}(c)\right|$, i.e., the value of the constant c for which the largest absolute value of the estimated conditional bias for the sample observations on the robust estimator is the lowest. In this case, the robust estimator is equal to:

$$\hat{T}^{\text{CB}}(X)(c^*) = \hat{T}^{\text{BHR}}(X) = \hat{T}(X) - \frac{\min_i \hat{B}_{1i}^{\hat{T}(X)} + \max_i \hat{B}_{1i}^{\hat{T}(X)}}{2}. \tag{3.8}$$

The Beaumont, Haziza and Ruiz-Gazen estimator is thus simple to implement. Compared to the Kokic and Bell method, it is more general because it is valid for all sampling designs and does not require any information outside the sample to be determined. In addition, it does not rely on any hypotheses about the variable of interest. The resulting estimator is robust under the sample design, while the Kokic and Bell estimator considers the sampling design and the distribution of the variable of interest. However, it is not designed to have the smallest mean square error, but to obtain an estimator on which the influence of each unit is limited, by minimizing the influence of the most influential unit.

The method has been extended to integrate more elements of the sample design and to adapt to certain situations. Favre-Martinoz et al. (2016) extended the method for a two-phase sampling design, which makes it possible to take non-response into account when it is assimilated to a second phase of Poisson drawing; Favre-Martinoz et al. (2015) proposed a method for ensuring the consistency of the robust estimators obtained when the parameters of interest are the totals of a variable in different domains included in one another.

4 Comparison of winsorization and conditional bias

In the previous section, we presented two types of methods for processing influential units applied to survey data:

- the Kokic and Bell winsorization, which aims to determine the winsorization thresholds that minimize the mean square error of the winsorized estimator under the sample design and the law of the variable of interest, which was initially conceived for a stratified simple random sampling design, but which we have extended to the case of Poisson sampling. The Kokic and Bell method, like its extension, is thus valid under hypotheses made about the law of the winsorized variable;
- the conditional bias method proposed by Beaumont, Haziza and Ruiz-Gazen, which potentially applies to all sampling designs and does not rely on any hypothesis on the law of the variable of interest; it aims to obtain the estimator for which the most influential unit has the least influence possible.

To compare the efficiency of these two methods, we performed two exercises:

1. simulations applied to the Poisson sampling;
2. a comparison on real data, applied to the data from the French labour cost and structure of earnings survey (ECMOSS).

4.1 Simulations in the case of a Poisson sampling

We performed a simulations study to examine the properties of the two robust estimators proposed in the context of a Poisson drawing. We carried out four scenarios to compare the efficiency of the two estimators, but also to study, in the case of the Kokic and Bell estimator, the model's robustness to a bad specification, i.e., to a modification between the learning model and the model that generated the sample data.

The simulation proceeds as follows:

- We consider $L = 1{,}000$ realizations of a certain model, which makes it possible to generate our learning set of $N = 5{,}000$ units;
- For each of these realizations, we calculate the optimal threshold K_l according to the method proposed in Section (2.2);
- Then we create $M = 10{,}000$ test sampling frames generated according to a (different) model on which we select a sample of expected size $n = 500$ following a Poisson drawing and calculate the robust estimator $\hat{\theta}_{(m)}$ with the threshold K_l calculated. As a comparison, we also calculate the robust estimator resulting from the method based on the conditional bias.

The inclusion probabilities, as well as the values of the variable, X were generated according to the following model:

$$U_i \sim \mathcal{L}\text{og} - \mathcal{N}(1; 1.1),$$

$$\pi_i = n \times \frac{U_i}{\sum_{i=1}^{N} U_i},$$

$$X_i = 2{,}000 \times \pi_i + \pi_i \epsilon_i + \delta_i V_i,$$

$$\epsilon_i \sim \mathcal{N}(0; 100), V_i \sim \mathcal{L}\text{og} - \mathcal{N}(\log(500); 1.2), \delta_i \sim \mathcal{B}(\omega),$$

where ω is the Bernoulli parameter, reflecting the proportion of influential values whose values are given in Table 4.1. The notation $\mathcal{L}\text{og} - \mathcal{N}$ denotes a log-normal distribution.

Table 4.1
Values of parameter ω used to generate populations

	Values of parameter ω	
Scenario	**Learning model**	**Test model**
1	0	0
2	0.01	0.01
3	0.01	0.1
4	0.1	0.01

Scenario 1 corresponds to the population model for which the extension of the Kokic and Bell method was developed in the Poisson case with $H = 1,$ but in which no or very few units are influential (the value of the parameter ω being fixed at 0). Scenario 2 corresponds to a situation in which this model applies, but in which a small proportion (1%) of units are influential. The model is, in scenarios 1 and 2, identical in the population used to calculate the threshold and the sample to which the threshold is applied.

In scenarios 3 and 4, the basic model is the same between the learning population and the sample, but the number of influential units varies between the two. In scenario 3, the learning population contains 10 times fewer influential units than the sample. Scenario 4 corresponds to the opposite scenario.

As a measure of the bias of an estimator $\hat{\theta}$ of a total $T,$ we calculated the relative Monte Carlo bias (as 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 $\hat{\theta}_{(m)}$ is the estimator $\hat{\theta}$ in the sample $m, m = 1, \ldots, M.$

We also calculated the relative efficiency of the robust estimators relative (RE) to the dilation estimator, $\hat{t}$:

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

Tables 4.2 and 4.3 represent the descriptive statistics associated with the $L = 1{,}000$ Monte Carlo values calculated according to the learning population considered.

Table 4.2
Descriptive statistics for scenarios 1 and 2 of the 1,000 simulations for $n = 500$

Statistic	Scenario							
	1				2			
Description	K&B		BHR		K&B		BHR	
	BR	RE	BR	RE	BR	RE	BR	RE
Min.	-0.2	100	-0.43	100	-9.0	1	-4.3	26
$Q1$	-0.1	100	-0.32	100	-2.9	35	-1.9	51
Median	0.0	100	-0.27	100	-1.8	50	-1.5	62
Mean	0.0	100	-0.27	100	-2.0	50	-1.6	62
$Q3$	0.0	100	-0.23	100	-1.0	64	-1.3	73
Max.	0.0	100	-0.14	100	-0.1	109	-0.6	91

Scenario 1 corresponds to a situation in which no or very few influential units are present in the population: the performance of the robust estimators is therefore identical to that of the usual Horvitz-Thompson estimator, with a relative bias very close to 0. Scenario 2 corresponds to the situation for which the extension of the Kokic and Bell method to the Poisson case was developed, with the introduction of influential units. The two robust estimators are more effective than the usual estimator, but the performance of the Kokic and Bell estimator in terms of the gain in mean square error is greater, with a median relative efficiency over the 1,000 simulations of 50%, compared to 62% for the conditional bias method. This result is expected given that the threshold of the Kokic and Bell method is explicitly determined to obtain the estimator with the smallest mean square error.

Table 4.3
Descriptive statistics for scenarios 3 and 4 on the 1,000 simulations for $n = 500$

Statistic	Scenario							
	3				4			
Description	K&B		BHR		K&B		BHR	
	BR	RE	BR	RE	BR	RE	BR	RE
Min.	-32.2	2	-7.8	27	-4.5	1	-4.3	26
$Q1$	-18.9	50	-5.1	59	-1.8	48	-1.9	51
Median	-13.9	82	-4.6	66	-1.5	70	-1.5	62
Mean	-14.2	89	-4.7	65	-1.5	68	-1.6	62
$Q3$	-9.3	138	-4.2	72	-1.2	91	-1.3	73
Max.	-0.01	537	-2.7	89	-0.6	100	-0.6	91

The performances of the two methods in scenario 3 are more contrasted. While over the set of simulations, the conditional bias method succeeds in reducing the mean square error of the estimators, with a minimum mean square error gain of 27%, the Kokic and Bell method deteriorates precision in more than a quarter of cases. The population on which the threshold was calculated contains, in this scenario, too few influential units compared to the sample for the calculated threshold to be effective.

In scenario 4, where the learning population contains more influential units than the sample, the performances of the two methods are of the same order of magnitude.

Therefore, these simulations show:

- that in the absence of influential units, the two robust estimation methods do not lead to a loss of estimation efficiency;
- that when applied in its hypotheses, the Kokic and Bell method leads to more accurate estimators than the conditional bias method;
- that the Kokic and Bell method is, however, sensitive to the data used to calculate thresholds; if these data are not generated according to the same model as the data to which the thresholds are applied, the method may lead to a loss of precision;
- that the conditional bias method always allows a gain in precision on these simulations, even if this gain is not optimal.

4.2 Application to the Survey on labour costs and wage structure

4.2.1 Presentation of the survey

The Survey on labour cost and structure of earnings (ECMOSS) is conducted by INSEE every year and harmonized at the European level. It is used to respond to European regulations on the production of statistics on both the cost of labour and structure of earnings which contribute to comparisons between European countries in terms of work time and costs.

ECMOSS is a survey of local business units (or establishments). It covers all sectors–both market and non-market–with the exception of agriculture, state administrations and certain activities (extraterritorial activities, embassies, consulates, activities of individuals acting as employers) and businesses with 10 or more employees. It covers establishments located in the metropolitan territory and in the overseas departments. Each sampled business answers two questionnaires: In the first, it must provide a certain amount of aggregated information on its workforce, payroll and a breakdown into its main elements (basic wages, bonuses, social contributions paid by the employer and by employees, etc.) and on the number of work hours of its employees; in the second, it details these elements for a randomly selected sample of its employees.

Given this survey method, the ECMOSS sample design has two stages:

- First stage: A sample of approximately 17,000 establishments is selected according to a stratified sampling design by sector, size of business, size of the establishment and geographical location;

- Second stage: In each establishment, a sample of employees is selected from the lists of employees reported by the establishment to social security organizations. The sampling design is drawn independently in each establishment and stratified by social category of the employees, distinguishing between managers and non-managers. The number of employees surveyed in each establishment varies according to its size, but is limited to 24 to prevent the survey from placing too much burden on businesses. In the end, around 150,000 employees are surveyed each year.

Each year, a certain number of establishments do not respond to the survey, and responding establishments do not systematically provide information for all their employees. Therefore, there is total non-response at each stage, which is handled by reweighting according to the homogeneous response group method. Next, the final sample of respondent employees, on which most operations are performed, is calibrated on the population of employees from the files of social security organizations.

Last, the sample of employees is obtained through a complex sample design, comprising two drawing stages (establishments and employees), with two drawing phases at each stage.

Given the very great variability of the establishments and their wage policy (both in terms of differences in the average levels of wages between establishments and differences in the dispersion of wages in the establishments), the sampling weights of the sampled employees are widely dispersed, and the accuracy of the estimators is sensitive to the influential values of the sample: for example, a very high level executive in a large business, or the athletes employed by a high-level sports club.

4.2.2 Parameter of interest

The main parameter of interest in the survey is the average hourly wage, calculated in different dissemination domains: sectors, sectors crossed with the employment size ranges of the businesses, and sectors crossed with the region in which the establishment is located. The estimators used later in our simulations are obtained by calculating the ratio of estimators by expansion of total remuneration over the total number of hours:

$$\hat{R}(D) = \frac{\sum_{i \in S \cap D} w_i e_i}{\sum_{i \in S \cap D} w_i h_i} \tag{4.1}$$

with S the sample of employees, D the domain of interest, e_i the annual remuneration of the employee i, h_i their annual hourly work volume and w_i the employee's estimation weight obtained by multiplying the selection probabilities and the response probabilities associated with each stage and phase of the sample design. Estimator (4.1) does not correspond to the estimator used in practice because it involves the initial weights corrected for non-response, while the estimator used in practice uses the calibrated weights. In the context of this article, the calibration phase was not taken into account, but it could have been using the classical residual technique and an additional degree of complexity which we deemed unnecessary to compare the two robust estimation methods.

4.2.3 How to adapt the processing methods for influential units to the ECMOSS sampling design

Estimator (4.1) is not the expansion estimator of a total, for which the previously described methods were designed. The problem can, however, be adapted to the framework of these two methods.

Indeed, an unbiased estimator of the variance of $\sum_{i \in S} w_i \hat{L}_i \left[\hat{R}(D)\right]$, with $\hat{L}_i \left[\hat{R}(D)\right] = \frac{e_i - \hat{R}(D) h_i}{\sum_{i \in S \cap D} w_i \, h_i} \mathbb{I}(i \in D)$ the estimated linearized variable of $\hat{R}(D)$, is also an asymptotically unbiased estimator of $V(\hat{R}(D))$. Thus, a robust estimator of the total of the linearized variable $\hat{L}_i \left[\hat{R}(D)\right]$ will also be a robust estimator for the influential units of $\hat{R}(D)$. Each method, applied to the estimated linearized variable, generates a winsorized value of this variable, denoted $\hat{L}_i^w \left[\hat{R}(D)\right]$. The effects of the processing of the influential units are then transferred to all other variables of interest of the survey through the estimation weight, by calculating a winsorized estimation weight:

$$w_i^w = w_i \frac{\hat{L}_i^w \left[\hat{R}(D)\right]}{\hat{L}_i \left[\hat{R}(D)\right]}.$$

We thus test the two methods of Kokic and Bell and Beaumont, Haziza and Ruiz-Gazen to estimate the total of $\hat{L}_i \left[\hat{R}(D)\right]$. However, each of the two methods requires adaptations to be applied to the sampling design and variables of interest of ECMOSS.

4.2.4 Adaptation of winsorization according to the Kokic and Bell method and its extension

The survey and the parameter of interest of the survey, even after linearization, do not fit with the framework of the Kokic and Bell method, whether it is the original method, or the extension presented previously. First, the ECMOSS sample is not selected using a stratified simple random survey or a Poisson sampling. Moreover, the variable to winsorize, the estimated linearized variable $\hat{L}_i \left[\hat{R}(D)\right]$, is not always positive. To apply the Kokic and Bell method to the ECMOSS case, we have made the following adaptations.

1. We apply the processing of the influential units as though the employees were directly selected by stratified simple random sampling (Poisson sampling for the extension) in strata defined by the sector, the number of employees of the business and the location of the employing establishment, by grouping certain modalities of this last variable to avoid generating pseudo-strata containing too few observations (we distinguish Île de France, the overseas departments and the rest of the country) and by the social category of the employee (distinguishing managers and non-managers). As the classical method acts as though the sample in each pseudo-stratum was selected by simple random sampling and thus all employees of the same pseudo-stratum have the same sampling weight, we do not consider the dispersion of the estimation weights in the pseudo-strata from the actual sampling design of the survey, and thus risk missing influential

units. In the case of the extension of the method, this dispersion of the weights is properly taken into account.

2. In each of these pseudo-strata, winsorization is not applied directly to the estimated linearized variable, but to a translated version of it.

More precisely, we define for each sampled employee:

$$\hat{T}_i\left[\hat{R}(D)\right] = \hat{L}_i\left[\hat{R}(D)\right] + \min{}_{j\in S}\hat{L}_j\left[\hat{R}(D)\right]$$

for which we calculate winsorization thresholds in the pseudo-strata according to the method initially proposed by Kokic and Bell and for its extension. We then deduce two sets of estimation weights used to estimate the average hourly wage in each domain of interest of the form:

$$w_i^w = w_i\,\frac{\hat{T}_i^w\left[\hat{R}(D)\right]}{\hat{T}_i\left[\hat{R}(D)\right]}.$$

We can thus only identify and process influential units with high values of the estimated linearized variable, i.e., employees whose hourly wage is higher than the average hourly wage in the domain of interest D. Units with low hourly wages cannot be identified by this method, but pose less problems for the accuracy of estimates, since the distribution of hourly wages is particularly skewed, with a very long tail on the right.

A final adaptation is necessary to adapt the method to the case of ECMOSS. This can only be used if observations of the variable of interest in each pseudo-stratum are available. Previous editions of the survey can be used. However, the tests performed to evaluate the efficiency of the Kokic and Bell method applied to the Annual Sectoral Surveys (see Deroyon, 2015) have shown that the use of responses to previous editions of the survey to calculate winsorization thresholds does not lead to the largest gains in accuracy. This is because the small number of observations available per stratum to calculate these thresholds are determined with too little precision, so that too many units can be winsorized, or conversely, influential units escape winsorization. We have chosen to use the auxiliary information available in the social security files on total remuneration paid annually to employees and their number of hours worked. These data are not those measured in the survey (in particular, the wages declared in the social security files form the tax base on which are calculated social contributions and tax contributions on wages, and not labour income paid to employees), but are strongly correlated with them.

4.2.5 Adaptation of Beaumont, Haziza and Ruiz-Gazen estimator

Because of its generality, the conditional bias method requires fewer adaptations to be applied to the ECMOSS. It can thus be applied directly to the variables of interest of the survey without the need to mobilize external data. However, calculating conditional biases while considering the whole sampling design is complex; therefore, for our simulations, we chose to apply the conditional bias methods as though the employees had been selected directly by a Poisson sampling, with the selection probabilities $1/w_i$, where w_i designates the estimation weight after correction for non-response of the employee i. The conditional bias used to identify influential units is therefore equal to:

$$B_{1i}\left\{\hat{L}_i\left[\hat{R}(D)\right]\right\} = (w_i - 1)\,\hat{L}_i\left[\hat{R}(D)\right].$$

With formula (3.8), the Beaumont, Haziza and Ruiz-Gazen estimator processes only a limited number of units, i.e., the observations with the lowest and highest conditional biases, for which all corresponding indicators define the sets $A_{\min}$ and $A_{\max}$:

$$A_{\min} = \operatorname{argmin}_{j\in S} B_{1j}\left\{\hat{L}_j\left[\hat{R}(D)\right]\right\}$$

$$A_{\max} = \operatorname{argmax}_{j\in S} B_{1j}\left\{\hat{L}_j\left[\hat{R}(D)\right]\right\}.$$

Thus, the processed estimation weight of the influential units is equal to:

$$w_i^{\mathrm{BHR}} = \begin{cases} \dfrac{(2\,|\,A_{\min}\,|-1)\,w_i + 1}{2\,|\,A_{\min}\,|} & \text{if } \; B_{1i}\left\{\hat{L}_i\left[\hat{R}(D)\right]\right\} \in A_{\min} \\[2ex] \dfrac{(2\,|\,A_{\max}\,|-1)\,w_i + 1}{2\,|\,A_{\max}\,|} & \text{if } \; B_{1i}\left\{\hat{L}_i\left[\hat{R}(D)\right]\right\} \in A_{\max} \\[2ex] w_i & \text{otherwise.} \end{cases}$$

where $|\,A_{\min}\,|$ and $|\,A_{\max}\,|$ respectively designate the cardinal of $A_{\min}$ and $A_{\max}$.

Compared to the Kokic and Bell method, the robust estimator based on conditional biases does not focus on influential units located in the right-hand part of the distribution of the estimated linearized variable, but identifies the influential units with very low and very high values for this variable. It also focuses on an a priori limited number of units, since only observations with the minimum and maximum conditional bias are modified.

4.2.6 Robust estimation on several domains of interest

As previously described, the domains of interest for the dissemination of the ECMOSS results are numerous. For the sake of simplicity of dissemination and to comply with the requirements of European regulations, each employee in the individual sample must have only one estimation weight, so adaptations are necessary:

- Robust estimators for several sets of domains of interest

  European regulations require the dissemination of results in sets of domains that intersect and are not included in one another, such as intersections of sectors and ranges of numbers of employees and crossings of sectors and regions. Sampled units may belong to more than one dissemination domain.

Ideally, the processing of influential units should be done in each domain of interest separately, so that a single observation may be associated with a different estimation weight for each dissemination domain to which it belongs. However, this solution is not possible for the reasons mentioned above.

Another solution is to apply both of the methods to the crossings of all the dissemination domains. The risk is then in applying the processes for estimators calculated on very small populations, for which many units are influential. Thus, for the estimation on real dissemination domains, too many units would be winsorized. The resulting estimators will be less precise than the robust estimators adapted to each domain, but potentially also less precise than the unprocessed estimators of the influential units because they are too biased.

- Robust estimators for all modalities of a domain of interest

  For a given set of domains (e.g., industry sectors), an observation can be identified as influential and processed for estimation on more than one dissemination domain, and thus have more than one final estimate weight. This is the case if the selection of an observation belonging to a dissemination domain has an influence on the selection of other units belonging to other dissemination domains (e.g., in the case of a stratified sampling, if the dissemination domains intersect the drawing strata).

  This situation is impossible for the Beaumont, Haziza and Ruiz-Gazen estimator, for which we assume the Poisson sampling design. However, this can happen for the Kokic and Bell method and its extensions as we implement them, because some dissemination domains do not consist of groupings of the pseudo-strata that we have formed. The situation is then the same as that exposed in the case of several sets of dissemination domains: the only way to maintain a unique estimation weight for each sampled unit is to apply the methods to pseudo-dissemination domains close to the real dissemination domains but made up of groupings of winsorization pseudo-strata. These pseudo-domains are in fact formed by intersections of sectors, a range of the number of employees of the businesses and the geographic location of the establishments (distinguishing only the three modalities specified above).

To evaluate the performance in terms of precision gains or losses of the methods defined above, we carried out a set of simulations based on the ECMOSS sampling design and data on wages and hours worked from the social security files, available for all employees and for which we are therefore able to compare the average hourly wages observed in the population with their various estimators. In these simulations, we compared the efficiency of the methods applied directly to each dissemination domain, which lead to the optimal results, and to the pseudo-dissemination domains defined above.

4.2.7 Simulations

The simulations are conducted in the social security files, from which the sample of employees is selected and which are available for all French employees. They are implemented as follows:

- the ECMOSS sampling design (including the selection of responding establishments and employees) is applied 5,000 times to produce 5,000 samples of employees, denoted S_m, $m = 1, \ldots, 5{,}000$;
- for each sample and each dissemination domain, we calculate the usual expansion estimator of $\hat{R}_m(D)$;
- the Kokic and Bell winsorization and conditional bias are applied to each sample according to different specifications:
    - the Kokic and Bell winsorization, classical or adapted to Poisson sampling, is applied only as though the real dissemination domains were the pseudo-dissemination domains defined above.
    - The Beaumont, Haziza and Ruiz-Gazen estimator is applied in each activity sector taken separately on the one hand, and on the other hand as though the pseudo-dissemination domains defined above were the real dissemination domains. For each dissemination domain, we can thus compare the performances of the conditional bias estimator applied in its optimal specification for this domain (producing an estimator $\hat{R}_m^{\mathrm{BHR}*}(D)$ of the average hourly wages in the domain) to the conditional bias method and the Kokic and Bell method (producing estimators $\hat{R}_m^{\mathrm{BHR}}(D)$, $\hat{R}_m^{\mathrm{KB}}(D)$ and $\hat{R}_m^{\mathrm{KB}_{\mathrm{poiss}}}(D)$ for the extension) applied according to specifications that are sub-optimal for this domain but simpler to implement.

For each robust estimator and each domain, we calculate the mean relative bias (RB) and the relative mean square error (RMSE) for all simulations by:

$$\mathrm{AB}\left[\hat{R}^{\mathrm{KB}}(D)\right] = \frac{\sum_{m=1}^{5{,}000}\left[\hat{R}_m^{\mathrm{KB}}(D) - R(D)\right]}{5{,}000}$$

$$\mathrm{AMSE}\left[\hat{R}^{\mathrm{KB}}(D)\right] = \frac{\sum_{m=1}^{5{,}000}\left[\hat{R}_m^{\mathrm{KB}}(D) - R(D)\right]^2}{5{,}000}$$

$$\mathrm{RB}\left[\hat{R}^{\mathrm{KB}}(D)\right] = 100\,\frac{\mathrm{AB}\left[\hat{R}^{\mathrm{KB}}(D)\right]}{R(D)}$$

$$\mathrm{RMSE}\left[\hat{R}^{\mathrm{KB}}(D)\right] = 100\,\frac{\mathrm{AMSE}\left[\hat{R}^{\mathrm{KB}}(D)\right]}{\mathrm{AMSE}\left[\hat{R}(D)\right]}$$

where, for example, for the classical Kokic and Bell method, $R(D)$ designates the average hourly wage observed in the social security files in the domain D and $\hat{R}(D)$ designates the usual expansion estimator of this parameter. Relative bias compares the bias of the robust estimator to the real value of the parameter. The relative mean square error measures the gain or loss of precision provided by the robust estimators relative to the usual estimator.

4.2.8 Simulation results

Among the different estimators tested in our simulations, the estimator obtained by applying the adaptation of the Kokic and Bell method to Poisson sampling is distinguished by extremely poor performances, summarized in Table 4.4. Application of the Kokic and Bell method extended to Poisson sampling for the ECMOSS results in a significant or even dramatic deterioration in the precision of the estimates.

Table 4.4
Statistics on the mean square error (MSE) ratio of the robust Kokic and Bell estimators applied to the Poisson sampling in the different domains of interest

Statistic	$\mathrm{RMSE}\left(\hat{R}_m^{\mathrm{KB}_{\mathrm{poiss}}}(D)\right)$		
	Domain		
	NACE*Workforce	NACE	NACE*NUTS
Min.	18	128	33
Mean	490	1,858	324
Max.	4,437	8,606	2,466

Figures 4.1, 4.2 and 4.3 focus on presenting the results of the conditional bias and classical Kokic and Bell methods, applied under the hypothesis of a stratified simple random sampling.

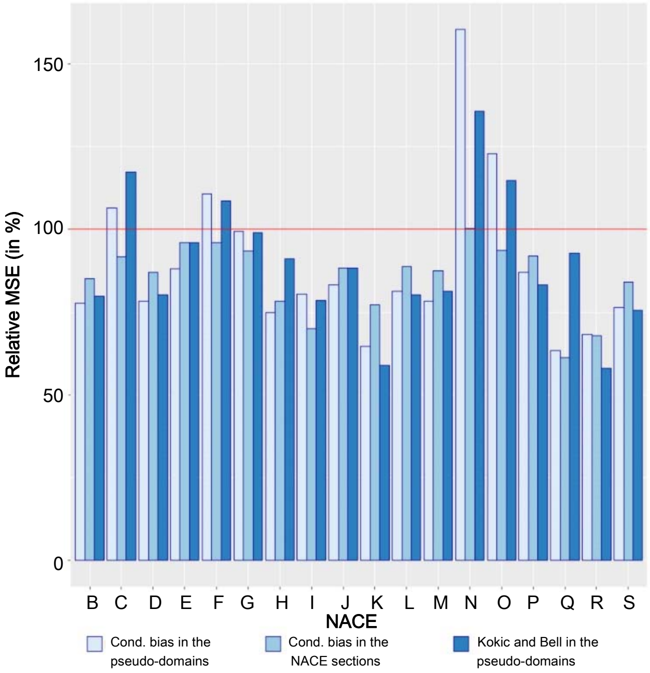


Figure 4.1 Relative mean square errors for the estimators of average hourly wage by sector.

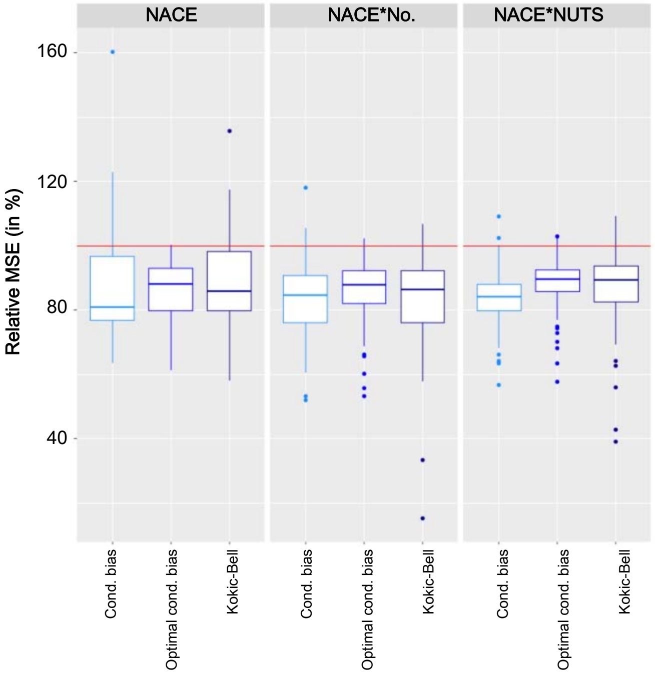


Figure 4.2 Distribution of relative mean square errors in each domain.

Figure 4.1 shows the relative mean square errors of the robust average hourly wage estimators in each section of the Statistical classification of economic activities in the European Community (NACE, a grouping of business sectors into 21 categories, of which 18 are in the ECMOSS field) and Figure 4.2 shows the distribution of relative mean square errors in each domain (among all sections, section crossings, and number of business employees, or crossings of sector and location of the establishment).

For almost all domains of interest, the robust estimators considered provide gains in precision over the usual expansion estimator. The domains in which the robust estimators have a higher error than the usual estimator are also those where the estimation variance is the lowest originally. The processes for influential units considered in these figures (conditional bias and classical Kokic and Bell method) are thus able to reduce estimation errors when necessary without causing too much loss of precision when the estimators are not affected by influential units.

The biases of the average hourly wage estimators in the sectors are low (see Figure 4.3), except in some domains where the sample size is small (A: Agriculture, forestry and fishing; K: Financial and insurance activities; R: Arts, entertainment and recreation). The same results are also observed for the other domains.

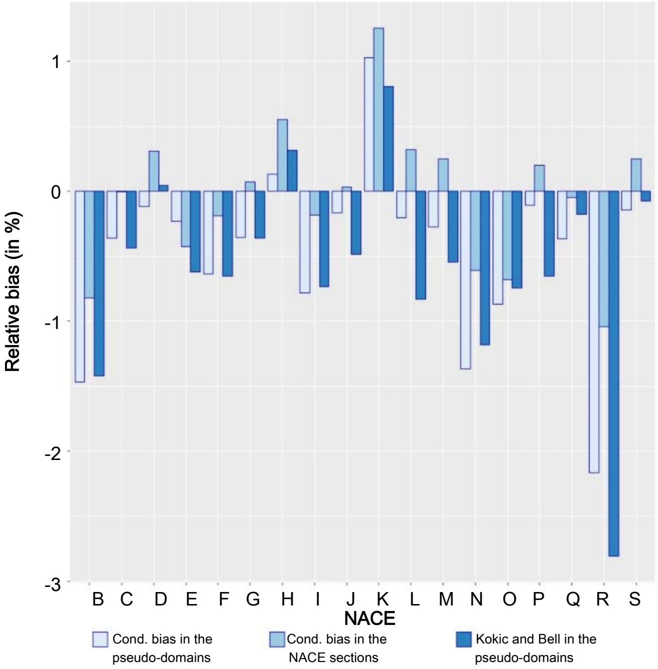


Figure 4.3 Relative biases for estimators of average hourly wage by sector.

The application of conditional bias methods adapted to each domain gives the best results for the estimation in the NACE sections, but not necessarily in the other dissemination domains. The NACE sections are much more aggregated than the pseudo-domains used for the identification of influential units, so the bias introduced by the processing of influential units is more significant in the cases where the application is made on pseudo-domains, compared to the optimal version applied directly to the NACE sections. In the other domains, the identification of influential units at a finer level than the real dissemination domain makes it possible to identify more influential units and thus substantially reduce the estimation variance, without introducing too much additional bias, when the domain used to identify the influential units and the real dissemination domains are close. Differences in how the sampling design is described to apply each of the two methods and the actual sampling design may explain why the use of the Beaumont, Haziza and Ruiz-Gazen robust estimator in each dissemination domain does not necessarily translate into greater precision gains.

The differences between the results obtained with the conditional bias and Kokic and Bell methods under the hypothesis of the stratified simple random sampling design are, however, small. Note however that, for the implementation of these simulations, we use the population data as observations of the additional interest variables not from the sample to calculate the winsorization thresholds in the Kokic and Bell method. Since we also evaluate the performance of the different estimators based on these data, the Kokic and Bell method is favoured a priori.

The extension of the Kokic and Bell method to Poisson sampling results in a significant deterioration in the precision of the estimators.

The discrepancies between the performances of the two implementations of the Kokic and Bell method are thus very high. However, these implementations are both based on two hypotheses:

- a hypothesis on the sampling design used to select the sample;
- a hypothesis on the distribution of the variable of interest in subpopulations U_h.

In both applications of the Kokic and Bell method, the first hypothesis is not respected. The violation of this hypothesis is, however, a priori more significant when we apply the Kokic and Bell method as though the sample had been selected by a stratified simple random sampling in pseudo-strata constructed ad-hoc, because in so doing we assume that the selection probabilities are identical in these pseudo-strata, which is not at all verified. The Kokic and Bell method applied as though the employees had been selected by Poisson sampling, for its part, considers real simple inclusion probabilities, but neglects the links between the indicators of belonging to the sample of different employees.

However, the population model postulated for the Kokic and Bell method extended to the Poisson case is not valid, since the simple inclusion probabilities are not proportional to the variable of interest considered. It is more complex to assess the validity of the population model used for the classical Kokic and Bell method; up to a point, it is still possible to consider that the results of the variable of interest in a pseudo-stratum are derived from the same law whose expectation and variance can be estimated by the mean and the empirical variance of the results of the variable of interest in the stratum.

Also, the performance differences of the two implementations of the Kokic and Bell method are complex to analyze. A first possible explanation is that the performances of the method are more sensitive to violations of the hypothesis on the law of the observations than to those on the form of the sampling design. This finding was shared by Fizzala (2017) in the case of an application of winsorization in the context of corporate profiling. In our ECMOSS simulations, we observe that the classical Kokic and Bell method, based on the hypothesis of stratified simple random samplings, gives very valid results despite the fact that this hypothesis is only partially respected. Future extensions of this work could consist of validating this explanation on the basis of simulations. Another explanation for these differences in performance may lie in the relationship between the two hypotheses in the case of the extension of Kokic and Bell to Poisson sampling. Indeed, while in the case of the classical Kokic and Bell method, the hypotheses on the sampling design and the law of the variable of interest in each stratum are unrelated, in the case of the Poisson sampling, the population model involves selection probabilities and therefore implies additional constraints on the sampling design. Therefore, the fact that the selection probabilities are not proportional to the variable of interest implies that, for the extension of the Kokic and Bell to Poisson sampling, the hypotheses on the sampling design and the population are simultaneously violated, which could explain this explosion of errors of the estimator.

However, the conditional bias and classical Kokic and Bell methods, whatever the configuration, seem to be able to identify influential units for the estimation of the parameters affected, and thus guarantee significant gains in precision even when they are applied in a setting that is remote from their original hypotheses and the actual sampling design of the survey.

Appendix

A Demonstrations of the formulas for the extension of the Kokic and Bell method in the case of a Poisson sampling

A.1 Calculation of the mean square error of the winsorized estimator

First, we will calculate

$$\begin{aligned} E_P\left\{\left[\hat{T}\left(\tilde{X}\right)-T\left(X\right)\right]^2\right\} &= E_P\left\{\left[\hat{T}\left(\tilde{X}\right)-T\left(\tilde{X}\right)\right]^2+\left[T\left(\tilde{X}\right)-T\left(X\right)\right]^2\right. \\ &\qquad \left.+\,2\left[\hat{T}\left(\tilde{X}\right)-T\left(\tilde{X}\right)\right]\left[T\left(\tilde{X}\right)-T\left(X\right)\right]\right\} \\ &= E_P\left\{\left[\hat{T}\left(\tilde{X}\right)-T\left(\tilde{X}\right)\right]^2\right\}+\left[T\left(\tilde{X}\right)-T\left(X\right)\right]^2 \end{aligned}$$

with $T\left(\tilde{X}\right)=E_P\left[\hat{T}\left(\tilde{X}\right)\right]=\sum_{h=1}^{H}\sum_{i\in U_h}\tilde{X}_{hi}$.

Furthermore,

$$E_P\left\{\left[\hat{T}\left(\tilde{X}\right)-T\left(\tilde{X}\right)\right]^2\right\}=\sum_{h=1}^{H}\sum_{i\in U_h}d_{hi}\left(1-\frac{1}{d_{hi}}\right)\tilde{X}_{hi}^2$$

finally:

$$E_P\left\{\left[\hat{T}\left(\tilde{X}\right)-T\left(X\right)\right]^2\right\}=\sum_{h=1}^{H}\sum_{i\in U_h}d_{hi}\left(1-\frac{1}{d_{hi}}\right)\tilde{X}_{hi}^2+\left[\sum_{h=1}^{H}\sum_{i\in U_h}\left(\tilde{X}_{hi}-X_{hi}\right)\right]^2. \qquad \text{(A.1)}$$

Assuming in each stratum that:

- $E_m\left(d_{hi}X_{hi}\right)=\mu_h$;
- $\mathrm{Var}_m\left(d_{hi}X_{hi}\right)=\sigma_h^2<+\infty$;
- and that $d_{hi}X_{hi}$ are independent and of density $g_h(x)>0$.

and noting that:

$$\begin{aligned} \tilde{X}_{hi} &= X_{hi}\left(1-J_{hi}\right)+J_{hi}\left[\frac{X_{hi}}{d_{hi}}+\left(1-\frac{1}{d_{hi}}\right)\frac{K_h}{d_{hi}}\right] \\ &= \frac{1}{d_{hi}}\left[d_{hi}X_{hi}+J_{hi}\left(1-\frac{1}{d_{hi}}\right)\left(K_h-d_{hi}X_{hi}\right)\right] \end{aligned}$$

and so that:

$$\tilde{X}_{hi}-X_{hi}=\frac{1}{d_{hi}}\left(1-\frac{1}{d_{hi}}\right)J_{hi}\left(K_h-d_{hi}X_{hi}\right), \qquad \text{(A.2)}$$

we obtain:

$$\begin{aligned}\tilde{X}_{hi}^{2} &= \frac{1}{d_{hi}^{2}}\left[d_{hi}^{2}X_{hi}^{2} + J_{hi}\left(1-\frac{1}{d_{hi}}\right)^{2}\left(K_{h}^{2} + d_{hi}^{2}X_{hi}^{2} - 2d_{hi}X_{hi}K_{h}\right)\right.\\ &\qquad \left. + 2\left(1-\frac{1}{d_{hi}}\right)\left(d_{hi}X_{hi}J_{hi}K_{h} - J_{hi}d_{hi}^{2}X_{hi}^{2}\right)\right], \end{aligned} \tag{A.3}$$

and that:

$$\begin{aligned} E_{m}\left\{\left[\sum_{h=1}^{H}\sum_{i\in U_{h}}\left(\tilde{X}_{hi} - X_{hi}\right)\right]^{2}\right\} &= \sum_{h=1}^{H}\sum_{i\in U_{h}} V_{m}\left(\tilde{X}_{hi} - X_{hi}\right)\\ &\quad + \left[\sum_{h=1}^{H}\sum_{i\in U_{h}} E_{m}\left(\tilde{X}_{hi} - X_{hi}\right)\right]^{2}\\ &= \sum_{h=1}^{H}\sum_{i\in U_{h}}\left\{E_{m}\left[\left(\tilde{X}_{hi} - X_{hi}\right)^{2}\right] - \left[E_{m}\left(\tilde{X}_{hi} - X_{hi}\right)\right]^{2}\right\}\\ &\quad + \left[\sum_{h=1}^{H}\sum_{i\in U_{h}} E_{m}\left(\tilde{X}_{hi} - X_{hi}\right)\right]^{2}. \end{aligned} \tag{A.4}$$

In the end, taking the expectation under the model of expression (A.1) and applying simplifications (A.2), (A.3), (A.4), we obtain, after some additional simplifications:

$$\begin{aligned} E_{m}E_{P}\left\{\left[\hat{\tilde{T}}\left(\tilde{X}\right) - T\left(X\right)\right]^{2}\right\} &= \sum_{h=1}^{H}\sum_{i\in U_{h}}\left(\frac{1}{d_{hi}}\right)\left(1-\frac{1}{d_{hi}}\right)\left\{\mu_{h}^{2} + \sigma_{h}^{2}\right.\\ &\quad + \left(1-\frac{1}{d_{hi}}\right)^{2}\left[K_{h}^{2}E_{m}\left(J_{hi}\right) + E_{m}\left(J_{hi}d_{hi}^{2}X_{hi}^{2}\right) - 2K_{h}E_{m}\left(J_{hi}d_{hi}X_{hi}\right)\right]\\ &\quad \left. + 2\left(1-\frac{1}{d_{hi}}\right)\left[K_{h}E_{m}\left(J_{hi}d_{hi}X_{hi}\right) - E_{m}\left(J_{hi}d_{hi}^{2}X_{hi}^{2}\right)\right]\right\}\\ &\quad + \sum_{h=1}^{H}\sum_{i\in U_{h}}\left(\frac{1}{d_{hi}}\right)^{2}\left(1-\frac{1}{d_{hi}}\right)^{2}\left\{K_{h}^{2}E_{m}\left(J_{hi}\right) + E_{m}\left(J_{hi}d_{hi}^{2}X_{hi}^{2}\right)\right.\\ &\quad \left. - 2K_{h}E_{m}\left(J_{hi}d_{hi}X_{hi}\right) + \left[K_{h}E_{m}\left(J_{hi}\right) - E_{m}\left(J_{hi}d_{hi}X_{hi}\right)\right]^{2}\right\}\\ &\quad + \left\{\sum_{h=1}^{H}\sum_{i\in U_{h}}\left(\frac{1}{d_{hi}}\right)\left(1-\frac{1}{d_{hi}}\right)\left[K_{h}E_{m}\left(J_{hi}\right) - E_{m}\left(J_{hi}d_{hi}X_{hi}\right)\right]\right\}^{2}. \end{aligned}$$

Given that the $d_{hi}X_{hi}$ are assumed to be independent and follow the same law within the strata, it is sufficient to consider a random variable Z_{h} that has the same law as one of the $d_{hi}X_{hi}$, , i.e., verifying:

- $E_{m}\left(Z_{h}\right) = \mu_{h}$;
- $\mathrm{Var}_{m}\left(Z_{h}\right) = \sigma_{h}^{2} < +\infty$;

- and that Z_h are independent and of density $g_h(x) > 0.$

Thus, we can also consider that a random variable $J_h = \mathbb{I}_{Z_h > K_h}$ to calculate the expectation with respect to the model of the winsorized indicator. The previous expression is rewritten:

$$\begin{aligned}
E_m E_P\left[\left(\hat{\tilde{T}}(\tilde{X}) - T(X)\right)^2\right] = &\sum_{h=1}^{H}\sum_{i\in U_h}\left(\frac{1}{d_{hi}}\right)\left(1-\frac{1}{d_{hi}}\right)\Bigg\{\mu_h^2 + \sigma_h^2 \\
&+\left(1-\frac{1}{d_{hi}}\right)^2\left[K_h^2 E_m(J_h) + E_m(J_h Z_h^2) - 2K_h E_m(J_h Z_h)\right] \\
&+ 2\left(1-\frac{1}{d_{hi}}\right)\left[K_h E_m(J_h Z_h) - E_m(J_h Z_h^2)\right]\Bigg\} \\
&+\sum_{h=1}^{H}\sum_{i\in U_h}\left(\frac{1}{d_{hi}}\right)^2\left(1-\frac{1}{d_{hi}}\right)^2\Big\{K_h^2 E_m(J_h) + E_m(J_h Z_h^2) - 2K_h E_m(J_h Z_h) \\
&- \left[K_h E_m(J_h) - E_m(J_h Z_h)\right]^2\Big\} \\
&+\left\{\sum_{h=1}^{H}\sum_{i\in U_h}\left(\frac{1}{d_{hi}}\right)\left(1-\frac{1}{d_{hi}}\right)\left[K_h E_m(J_h) - E_m(J_h Z_h)\right]\right\}^2.
\end{aligned}$$

A.2 Search for thresholds to minimize the MSE

To determine the value of the thresholds K_h leading to the optimum of $E_m E_P\left\{\left[\hat{T}(\tilde{X}) - T(X)\right]^2\right\},$ we use the same property as Kokic and Bell in their demonstration, i.e., that:

$$E_m\left(Z_h^p J_h\right) = \int_{K_h}^{+\infty} t_h^p\, g_h(t)\,dt,$$

and so that

$$\frac{\partial}{\partial K_h} E_m\left(Z_h^p J_h\right) = -K_h^p g_h(K_h).$$

By deriving relative to $K_h,$ and after simplification, we obtain that:

$$\begin{aligned}
\frac{\partial}{\partial K_h} E_m E_P\left\{\left[\hat{\tilde{T}}(\tilde{X}) - T(X)\right]^2\right\} = &\ 2B \times A_h E_m(J_h) \\
&+ 2C_h\left\{\left[K_h E_m(J_h) - E_m(J_h Z_h)\right]\left[1 - E_m(J_h)\right]\right\} \\
&+ 2D_h\left[K_h E_m(J_h) - E_m(J_h Z_h)\right] + 2F_h E_m(J_h Z_h) \qquad \text{(A.5)}
\end{aligned}$$

where

- $B = \sum_{h=1}^{H}\sum_{i\in U_h}\left(\frac{1}{d_{hi}}\right)\left(1-\frac{1}{d_{hi}}\right)\left[K_h E_m(J_h) - E_m(J_h Z_h)\right],$
- $A_h = \sum_{i\in U_h}\left(\frac{1}{d_{hi}}\right)\left(1-\frac{1}{d_{hi}}\right),$
- $C_h = \sum_{i\in U_h}\left(\frac{1}{d_{hi}}\right)^2\left(1-\frac{1}{d_{hi}}\right)^2,$

- $D_h = \sum_{i \in U_h} \left(\frac{1}{d_{hi}}\right)\left(1 - \frac{1}{d_{hi}}\right)^3,$
- $F_h = \sum_{i \in U_h} \left(\frac{1}{d_{hi}}\right)\left(1 - \frac{1}{d_{hi}}\right)^2.$

Equation (A.5) is reduced to:

$$\frac{\partial}{\partial K_h} E_m E_P \left[\left(\hat{\tilde{T}}\left(\tilde{X}\right) - T\left(X\right)\right)^2\right] = 0$$

$$\Leftrightarrow$$

$$A_h \times B \times E_m\left(J_h\right) + \left(C_h + D_h\right) K_h E_m\left(J_h\right)$$
$$- C_h E_m\left(J_h\right)\left[K_h E_m\left(J_h\right) - E_m\left(J_h Z_h\right)\right] + \left(F_h - C_h - D_h\right) E_m\left(J_h Z_h\right) = 0.$$

Finally, by noting that $\left(F_h - C_h - D_h\right) = 0$ and assuming that $E_m\left(J_h\right) > 0,$ we obtain that the threshold K_h minimizing the MSE verifies the equation:

$$A_h \times B + \left(C_h + D_h\right) K_h - C_h\left[K_h E_m\left(J_h\right) - E_m\left(J_h Z_h\right)\right] = 0$$

which is reduced further to

$$B + \frac{\left(C_h + D_h\right)}{A_h} K_h = \frac{C_h}{A_h}\left[K_h E_m\left(J_h\right) - E_m\left(J_h Z_h\right)\right].$$

It remains to be shown that $\frac{C_h\left[K_h E_m\left(J_h\right) - E_m\left(J_h Z_h\right)\right]}{A_h B}$ tends toward zero when $n \rightarrow \infty.$ However,

$$\frac{C_h \left| K_h E_m\left(J_h\right) - E_m\left(J_h Z_h\right)\right|}{\left| A_h B\right|} = \frac{C_h \left| K_h E_m\left(J_h\right) - E_m\left(J_h Z_h\right)\right|}{\left| A_h\right| \sum_{l=1}^{H} \left| A_l\right| \left| K_l E_m\left(J_l\right) - E_m\left(J_l Z_l\right)\right|}$$

and according to hypothesis (2.8) relating to inclusion probabilities, we have that, $\forall h = 1, \ldots, H, \forall i \in U_h$ $d_{hi} > 1.$ Which implies $A_h > 0,$ and thus:

$$\begin{aligned} \frac{\left| C_h\right| \left| K_h E_m\left(J_h\right) - E_m\left(J_h Z_h\right)\right|}{\left| A_h B\right|} &\leq \frac{C_h}{A_h^2} \\ &\leq \frac{\sum_{i \in U_h} \left(\frac{1}{d_{hi}}\right)^2 \left(1 - \frac{1}{d_{hi}}\right)^2}{\left[\sum_{i \in U_h} \left(\frac{1}{d_{hi}}\right)\left(1 - \frac{1}{d_{hi}}\right)\right]^2} \\ &\leq \frac{1}{\left[\sum_{i \in U_h} \left(\frac{1}{d_{hi}}\right)\left(1 - \frac{1}{d_{hi}}\right)\right]}. \end{aligned}$$

However, it is possible to demonstrate from hypothesis (2.8) that $\left[\sum_{i \in U_h} \left(\frac{1}{d_{hi}}\right)\left(1 - \frac{1}{d_{hi}}\right)\right]^{-1} = O\left(\frac{1}{N_h}\right).$ Thus: $\frac{C_h\left[K_h E_m\left(J_h\right) - E_m\left(J_h Z_h\right)\right]}{A_h B}$ tends toward zero when $n \rightarrow \infty.$

K_h is thus equivalent in each stratum to $-\frac{A_h}{\left(C_h + D_h\right)} B,$ when the size of the population and the sample tend toward infinity.

References

Chambers, R. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81, 1063-1069.

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

Clark, R.G. (1995). Winsorization methods in sample surveys. Master's thesis, Department of Statistics, Australian National University.

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

Demoly, E., Fizzala, A. and Gros, E. (2014). Méthodes et pratiques des enquêtes entreprises à l'Insee. *Journal de la Société Française de Statistique*, 155-4.

Deroyon, T. (2015). Traitement des observations atypiques d'une enquête par winsorisation : application aux Enquêtes Sectorielles Annuelles. *Actes des Journées de Méthodologie Statistique*.

Fizzala, A. (2017). *Adaptations of Winsorization Caused by Profiling - An Example Based on the French SBS Survey*. European Establishment Survey Workshop, Southampton.

Favre-Martinoz, C., Haziza, D. and Beaumont, J.-F. (2015). A method of determining the winsorization threshold, with an application to domain estimation. *Survey Methodology*, 41, 1, 57-77. Paper available at https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2015001/article/14199-eng.pdf.

Favre-Martinoz, C., Haziza, D. and Beaumont, J.-F. (2016). Robust inference in two-phase sampling designs with application to unit nonresponse. *Scandinavian Journal of Statistics*, 43, 1019-1034.

Kokic, P.N., and Bell, P.A. (1994). Optimal winsorizing cut-offs for a stratified finite population estimation. *Journal of Official Statistics*, 10-4, 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, 923-968.

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

Rivest, L.-P., and Hurtubise, D. (1995). On searls' winsorized mean for skewed populations. *Survey Methodology*, 21, 2, 107-116. Paper available at https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1995002/article/14399-eng.pdf.

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, 229-234.

Survey Methodology, December 2018
Vol. 44, No. 2, pp. 339-349
Statistics Canada, Catalogue No. 12-001-X

Criteria for choosing between calibration weighting and survey weighting

Mohammed El Haj Tirari and Boutaina Hdioud[1]

Abstract

Based on auxiliary information, calibration is often used to improve the precision of estimates. However, calibration weighting may not be appropriate for all variables of interest of the survey, particularly those not related to the auxiliary variables used in calibration. In this paper, we propose a criterion to assess, for any variable of interest, the impact of calibration weighting on the precision of the estimated total. This criterion can be used to decide on the weights associated with each survey variable of interest and determine the variables for which calibration weighting is appropriate.

Key Words: Estimation of a total; calibration estimator; superpopulation model; model-based approach; weighting impact.

1 Introduction

When estimating population parameters, adjustment techniques are often used to reduce variance or correct non-response. When there is auxiliary information, calibration is an adjustment technique often used in practice. The weight of the calibration estimator is used to adjust the sample so that it reflects the known population totals for a set of auxiliary variables (Deville and Särndal, 1992). The improved accuracy by the calibration estimator depends on the auxiliary variables used in calibration. The variance of the calibration estimator is low when the calibration variables are strongly linked to the variable of interest.

In practice, once the calibration weights are calculated, they replace the survey weights for the production of parameter estimates of all survey variables of interest. However, using calibration weighting can lead to an increase in the mean square error (MSE) for some variables of interest, particularly those not linked to calibration variables. Therefore, calibration weights cannot be used systematically to estimate population parameters for any variable of interest, particularly in the case of multi-purpose surveys covering different subjects. That is why it is necessary to develop a criterion to assess the impact of calibration weighting on the precision of estimates for each variable of interest.

To develop this type of criterion, we can refer to a comparison of the precision of calibration estimators with the Horvitz-Thompson (HT) estimator. Several inferential approaches can be used to measure the precision of these estimators. In this paper, we will consider a sample design- and model-based approach. This approach was chosen because it is the only one with which we can develop a measurement of the MSE of the calibration estimator in order to account for bias due to the use of calibration weights, as well as variance, which depends on the quality of the model. In other approaches (design-based or model-assisted), it is extremely difficult to calculate the MSE of the calibration estimator, and the estimates do not take into account the bias introduced by the use of calibration weights.

[1] Mohammed El Haj Tirari, Institut National de Statistique et d'Économie Appliquée, Rabat, Morocco. E-mail: mtirari@hotmail.fr; Boutaina Hdioud, École Nationale Supérieure d'Informatique et d'Analyse des Systèmes, Rabat, Morocco. E-mail: hdioud.boutaina@hotmail.fr.

Using the design- and model-based approach allows us to develop a criterion with the advantage of approaching a situation where the loss in bias increase for the calibration estimator exceeds the gain in the reduction of variance obtained when there is a link between the variable of interest and the calibration variables. This is a case where the calibration estimator must not be used.

In this paper, we propose a new criterion that measures the impact of using calibration weighting. The proposed criterion takes into account the degree of the existing link between the variable of interest and the calibration variables. Furthermore, it is simple to calculate for each survey variable of interest so that the best sets of weights to use can be identified.

It should be noted that the impact of using calibration weights was studied previously, but only in the context of measuring the design effect (Deff) used to assess the relative increase or decrease in the variance of an estimator compared with simple random sampling. For example, in the model-assisted approach, Henry and Valliant (2015) proposed a Deff measurement that translated the joint impact of an unequal probability sample design and an adjustment of sampling weights compared with simple random sampling.

Following the introduction, which identifies the issue examined in this paper, Section 2 presents the inferential approach adopted in this paper and the criterion used to measure the precision of estimators, while determining its expression for a calibration estimator and an HT estimator. In Section 3, we present the proposed new criterion for assessing the impact of using calibration weights. Section 4 evaluates the proposed criterion using simulations. The purpose of this evaluation is to verify that this criterion identifies situations where a set of calibration weights should be used. In Section 5, we conclude with a discussion of the advantages of the proposed criterion.

2 Estimator of a variable of interest total

$U = \{1, \ldots, N\}$ for a population size N from which sample s of size n is selected based on survey design $p(s)$. S designates a random variable such as $p(s) = P(S = s)$, and π_k and π_{kl} respectively designate the first and second probabilities of inclusion in survey design $p(s)$. We are interested in a variable of interest $Y = (y_1, \ldots, y_k, \ldots, y_N)'$, with the objective of estimating its total $t_y = \sum_{k \in U} y_k$. To do that, we consider the category of linear estimators $\hat{t}_{yw} = \sum_{k \in S} w_{kS} y_k$ where w_{kS} are the weights that can depend on sample S and the auxiliary variables available. The basic weights used are the sampling weights generated by $d_k = 1/\pi_k$. They correspond to the Horvitz-Thompson estimator $\hat{t}_{y\pi}$ (1952).

It is assumed that we have p auxiliary variables $X_1, \ldots, X_p$, for which the values may be represented by vectors $\mathbf{x}_k = (x_{k1}, \ldots, x_{kp})'$ and for which the vector of their totals $t_{\mathbf{x}} = \sum_{k \in U} \mathbf{x}_k$ is known. The category of calibration estimators is defined by $\hat{t}_{yC} = \sum_{k \in S} w_{kS,C} y_k$ where $w_{kS,C}$, referred to as calibration weights, verify the calibration equation given by

$$\sum_{k \in S} w_{kS,C} \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k. \tag{2.1}$$

Calibration helps to reduce the variance of a total estimator, particularly for variables of interest that are linked to the auxiliary variables used in calibration. However, calibration results in an estimator with a bias other than zero. That is why the calibration weights are determined so that they are as close as possible to the sampling weights in order to manage bias.

2.1 Precision of a linear total estimator

In order to measure the precision of a linear total estimator, we will consider the design and model-based approach. In addition to the design distribution, this approach consists of assuming that values $y_1, \ldots, y_k, \ldots, y_N$ for the variable of interest Y are the product of a random vector $(Y_1, \ldots, Y_k, \ldots, Y_N)'$ whose joint probability distribution is given by the *Superpopulation* model ξ defined by:

$$Y_k = \mathbf{x}'_k \boldsymbol{\beta} + \varepsilon_k \tag{2.2}$$

with

$$E_\xi(\varepsilon_k) = 0, \quad \mathrm{Var}_\xi(\varepsilon_k) = \sigma_k^2 \quad \text{and} \quad \mathrm{Cov}_\xi(\varepsilon_k, \varepsilon_l) = 0$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$, $\sigma_k^2\ (k \in U)$ are unknown parameters. E_ξ, Var_ξ and Cov_ξ represent respectively the expectation, variance and covariance for the model. Vector estimator $\boldsymbol{\beta}$ for the regression coefficients is produced by

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'_S \boldsymbol{\Pi}_S^{-1} \mathbf{V}_S^{-1} \mathbf{X}_S\right)^{-1} \mathbf{X}'_S \boldsymbol{\Pi}_S^{-1} \mathbf{V}_S^{-1} \mathbf{Y}_S$$

where $\mathbf{X}'_S$ is the matrix of $\mathbf{x}'_k$ values for $k \in S$, $\boldsymbol{\Pi}_S = \mathrm{diag}(\pi_k)_{k \in S}$ and $\mathbf{V}_S = \mathrm{diag}(\sigma_k^2)_{k \in S}$. Under the the design and model-based approach, the criterion used to measure the precision of a linear total estimator is

$$\mathrm{MSE}_{p\xi}(\hat{t}_{yw}) = E_p E_\xi (\hat{t}_{yw} - t_y)^2 \tag{2.3}$$

which corresponds to the mean square error (MSE) for the design and model, also referred to as the *anticipated mean square error* (AMSE). This is based on the assumption that the design is not informative. We can then show that the AMSE for linear estimator $\hat{t}_{yw}$ is (Nedyalkova and Tillé, 2008):

$$\mathrm{MSE}_{p\xi}(\hat{t}_{yw}) = E_p \left(\sum_{k \in s} w_{kS} \mathbf{x}'_k \boldsymbol{\beta} - \sum_{k \in U} \mathbf{x}'_k \boldsymbol{\beta} \right)^2 + \sum_{k \in U} \sigma_k^2 \left[\mathrm{var}_p(w_{kS} I_k) + (R_{kS} - 1)^2 \right] \tag{2.4}$$

where

$$R_{kS} = \frac{E(w_{kS} \mid I_k = 1)}{d_k}$$

with $d_k = 1/\pi_k$ (sampling weight) and $I_k = 1$ for $k \in S$ and $I_k = 0$ otherwise. Ratio R_{kS} equals 1 when linear estimator $\hat{t}_{yw}$ is unbiased according to the design.

2.2 AMSE for the calibration estimator

For the calibration estimator, verifying the calibration equation renders it unbiased under the model:

$$E_{\xi}\left(\hat{t}_{yC} - t_y\right) = \sum_{k\in S} w_{kS,C}\mathbf{x}'_k\boldsymbol{\beta} - \sum_{k\in U}\mathbf{x}'_k\boldsymbol{\beta} = 0.$$

Consequently, the AMSE is expressed as:

$$\begin{aligned}\text{MSE}_{p\xi}\left(\hat{t}_{yC}\right) &= \sum_{k\in U}\sigma_k^2\left[\text{var}_p\left(w_{kS,C}I_k\right) + \left(R_k - 1\right)^2\right]\\ &= \sum_{k\in U}\sigma_k^2\left[\frac{V_k}{d_k} + R_k^2\left(d_k - 1\right) + \left(R_k - 1\right)^2\right] \qquad (2.5)\end{aligned}$$

where $V_k = \text{var}_p\left(w_{kS,C} \mid I_k = 1\right)$ and $R_k = E_p\left(w_{kS,C} \mid I_k = 1\right)/d_k$.

Giving

$$\begin{aligned}\text{var}_p\left(w_{kS,C}I_k\right) &= E_p\left[\text{var}_p\left(w_{kS,C}I_k \mid I_k\right)\right] + \text{var}_p\left[E_p\left(w_{kS,C}I_k \mid I_k\right)\right]\\ &= \pi_k\,\text{var}_p\left(w_{kS,C} \mid I_k = 1\right) + \pi_k\left[E_p\left(w_{kS,C} \mid I_k = 1\right)\right]^2 - \left[E_p\left(w_{kS,C}I_k\right)\right]^2\\ &= \frac{V_k}{d_k} + R_k^2\left(d_k - 1\right). \qquad (2.6)\end{aligned}$$

Note that the expression (2.5) of $\text{MSE}_{p\xi}\left(\hat{t}_{yC}\right)$ makes it possible to underscore the two criteria that determine the accuracy of calibration estimator $\hat{t}_{yC}$. The first corresponds to *Superpopulation model* ξ through its residual variance σ_k^2, which decreases when the variable of interest and the calibration variables are correlated (variance reduction $\hat{t}_{yC}$). The second criterion is represented by weight ratios R_k, which become important when the calibration weights are very different from the sampling weights (bias increase $\hat{t}_{yC}$).

2.3 AMSE for the HT estimator

In order to develop our criterion for choosing between calibration weighting and sample weighting, we need to determine the expression of the AMSE for the HT estimator. Since the latter is unbiased under the design $\left(R_{kS} = 1\right)$, its AMSE is given by:

$$\begin{aligned}\text{MSE}_{p\xi}\left(\hat{t}_{y\pi}\right) &= \text{var}_p\left(\sum_{k\in s} d_k\mathbf{x}'_k\boldsymbol{\beta}\right) + \sum_{k\in U}\sigma_k^2 d_k\left(1 - \pi_k\right)\\ &= \sum_{k\in U}\sum_{l\in U}\left(\pi_{kl} - \pi_k\pi_l\right)d_k\mathbf{x}'_k\boldsymbol{\beta}d_l\mathbf{x}'_l\boldsymbol{\beta} + \sum_{k\in U}\sigma_k^2 d_k\left(1 - \pi_k\right). \qquad (2.7)\end{aligned}$$

It should be noted that the expression of the AMSE for $\hat{t}_{y\pi}$ depends on probabilities π_{kl}, which are generally unknown and difficult to calculate for unequal probability sampling designs. Several approximations for these probabilities have been proposed in literature, enabling us to obtain several possible estimators for the variance of the HT estimator. However, Matei and Tillé (2005) showed, through a series of simulations, that these estimators are almost equivalent and allow us to effectively estimate the exact expression of the variance under design $\hat{t}_{y\pi}$.

An approximation of $\operatorname{var}_p \left(\sum_{k \in s} d_k \mathbf{x}'_k \boldsymbol{\beta}\right)$ can be obtained by considering the one proposed by Hájek (1981) for the variance of the HT estimator, produced by:

$$V_{\text{Approx}} = \sum_{k \in U} c_k \left(d_k \mathbf{x}'_k \boldsymbol{\beta}\right)^2 - \frac{1}{h}\left(\sum_{k \in U} c_k d_k \mathbf{x}'_k \boldsymbol{\beta}\right)^2 \tag{2.8}$$

where $h = \sum_{k \in U} c_k$ and $c_k = N\pi_k (1 - \pi_k)/(N - 1)$. The latter is obtained from the following approximation of probabilities π_{kl} (see Deville and Tillé, 2005; Tirari, 2003):

$$\pi_{kl} - \pi_k \pi_l \approx \begin{cases} c_k - \dfrac{c_k^2}{h} & \text{if } k = l \\ -\dfrac{c_k c_l}{h} & \text{if } k \neq l. \end{cases} \tag{2.9}$$

Consequently, the AMSE for $\hat{t}_{y\pi}$ can be approximated by:

$$\widetilde{\text{MSE}}_{p\xi}\left(\hat{t}_{y\pi}\right) = V_{\text{Approx}} + \sum_{k \in U} \sigma_k^2 d_k (1 - \pi_k). \tag{2.10}$$

It should be noted that for simple designs, such as Poisson design or simple stratified random design, joint probability can be calculated precisely without the need for an approximation. In the next section, we will be basing calibration and HT estimators on the AMSE to develop a new *measurement* of the impact of using calibration weights.

3 Proposed criterion for measuring the impact of using calibration weights

Calibration weights are used to improve the precision of estimates for survey parameters of interest. This improvement depends largely on how strongly the variable of interest is linked to the calibration variables. To assess the impact of using calibration weights, we can compare the AMSE for estimators $\hat{t}_{yC}$ and $\hat{t}_{y\pi}$ given respectively by (2.5) and (2.10). The impact of using calibration weights can then be measured through the following criterion:

$$\text{Weff} = \frac{\sum_{k \in U} \sigma_k^2 \left[\frac{V_k}{d_k} + R_k^2 (d_k - 1) + (R_k - 1)^2\right]}{V_{\text{Approx}} + \sum_{k \in U} \sigma_k^2 d_k (1 - \pi_k)} \tag{3.1}$$

where calibration weights are chosen in cases where the Weff value is less than 1. Note that the Weff expression (3.1) depends on the population and must be estimated. Furthermore, for any $k \in U, V_k$ represents the variance of calibration weight $w_{kS,C}$, considering the s set of samples containing unit k. Variance V_k is generally not zero since the $w_{kS,C}$ weights depend on the calibration variables and the s

sample selected. In order to take variance V_k into account in measuring the impact of using calibration weights $w_{kS,C}$, we propose estimating the quantity

$$V_w = \sum_{k \in U} \sigma_k^2 \frac{V_k}{d_k} \tag{3.2}$$

by

$$\hat{V}_w = \sum_{k \in S} \hat{\sigma}_k^2 \left(w_{kS,C} - d_k\right)^2 \tag{3.3}$$

where $\hat{\sigma}_k^2$ is the White estimator for σ_k^2 defined by $n\hat{\varepsilon}_k^2 / (n - p)$ with $\hat{\varepsilon}_k = Y_k - \mathbf{x}_k' \hat{\boldsymbol{\beta}}.$ The estimator (3.3) is obtained by replacing V_k by $\left(w_{kS,C} - d_k\right)^2$, which can be viewed as a first-order approximation of V_k. For any unit $k \in U$, the use of calibration produces weight $w_{kS,C}$, which varies from one sample to another, but for which the design-based expectation can be approximated by sampling weight d_k. The simulations discussed in Section 4 show that $\hat{V}_w$ is a good V_w estimator since it helps to deduct an effective estimator of the Weff criterion. The Weff criterion that we propose for choosing between calibration weights $w_{kS,C}$ and sampling weights d_k can be estimated by

$$\widehat{\text{Weff}}_S = \frac{\sum_{k \in S} d_k \hat{\sigma}_k^2 \left[\frac{\left(w_{kS,C} - d_k\right)^2}{d_k} + \hat{R}_{kS}^2 \left(d_k - 1\right) + \left(\hat{R}_{kS} - 1\right)^2\right]}{\hat{V}_{\text{Approx},S} + \sum_{k \in S} d_k \hat{\sigma}_k^2 \left(d_k - 1\right)} \tag{3.4}$$

where $\hat{R}_{kS} = w_{kS} / d_k$ and $\hat{V}_{\text{Approx},S}$ is an estimator for $\text{var}_p\left(\sum_{k \in S} d_k \mathbf{x}_k' \boldsymbol{\beta}\right)$ resulting from the approximation (2.8). It is produced by:

$$\hat{V}_{\text{Approx},S} = \sum_{k \in S} \tilde{c}_k \left(d_k \mathbf{x}_k' \hat{\boldsymbol{\beta}}\right)^2 - \frac{1}{\hat{h}} \left(\sum_{k \in S} \tilde{c}_k d_k \mathbf{x}_k' \hat{\boldsymbol{\beta}}\right)^2 \tag{3.5}$$

with $\tilde{c}_k = n\left(1 - \pi_k\right) / (n - 1)$ and $\hat{h} = \sum_{k \in S} \tilde{c}_k$. The proposed $\widehat{\text{Weff}}_S$ criterion has the benefit of considering bias due to the use of calibration weights, through $\hat{R}_{kS}$, as well as the quality of the linear regression model representing the link between the variable of interest and the calibration variables, through variance $\hat{\sigma}_k^2$. For some survey designs, the weighting traditionally used for estimates effectively leads to an unbiased estimator for the design, but it is not necessarily the HT estimator. This is the case, for example, with a two-stage design where the second stage design depends on the sample from the first stage and the weighting used is the product of the sampling weights for each stage. It is important to note that the $\widehat{\text{Weff}}_S$ criterion proposed in this paper is not linked to the HT estimator, since it enables us to compare the calibration estimator with any other estimator using the sampling weights once it is unbiased.

4 Simulation study

In order to evaluate the $\widehat{\text{Weff}}_S$ criterion (3.4), so that we can determine whether to use calibration weights or sampling weights, we conducted a series of simulations using data observed for a population of

5,800 cottage-industry units. We considered six calibration variables, from which several variables of interest Y_i were generated, with consideration for linear regression models, while accounting for the strength of the link between the variables of interest and the calibration variables through the choice of residual variance in the regression models. Furthermore, to study the impact of the heteroskedasticity of the model residuals on the results obtained for criterion $\widehat{\mathrm{Weff}}_s$, we also considered the case where the variables of interest are generated using models with heteroskedastic residuals.

For the purposes of these simulations, we selected 10,000 samples using a simple random sampling design (SRSD), with three sample sizes: 100, 200 and 400 cottage-industry units, to study the impact of the sample size on the results obtained. Across the 10,000 samples selected, we calculated the following indicators:

- $\mathrm{MSE}_{\mathrm{Cal}}$: the AMSE for the calibration estimator, the expression of which is given by (2.5) and where $E\left(w_{kS,C} \mid I_k = 1\right)$ and V_k are determined respectively by the mean and the variance of weights $w_{kS,C}$ considering all of the selected samples containing unit k.
- $\widetilde{\mathrm{MSE}}_{\mathrm{HT}}$: approximation (2.10) of the AMSE for the HT estimator. $\widetilde{\mathrm{MSE}}_{\mathrm{HT}}$ corresponds to $\mathrm{MSE}_{\mathrm{HT}}$ (AMSE (2.7) for the HT estimator) that we were able to calculate in these simulations since the samples were selected using SRSD.
- Weff: the theoretical value of the Weff calculated using (3.1) and defined by the ratio of $\mathrm{MSE}_{\mathrm{Cal}}$ and $\widetilde{\mathrm{MSE}}_{\mathrm{HT}}$.
- $\overline{\widehat{\mathrm{MSE}}}_{\mathrm{Cal}}$: the simulation mean for the $\widehat{\mathrm{MSE}}_{\mathrm{Cal}}$ estimator of $\mathrm{MSE}_{\mathrm{Cal}}$ where

$$\overline{\widehat{\mathrm{MSE}}}_{\mathrm{Cal}} = \frac{1}{10{,}000} \sum_{s=1}^{10{,}000} \left(\sum_{k \in s} d_k \hat{\sigma}_k^2 \left[\frac{\left(w_{ks,C} - d_k\right)^2}{d_k} + \hat{R}_{ks}^2 \left(d_k - 1\right) + \left(\hat{R}_{ks} - 1\right)^2 \right] \right).$$

- $\overline{\widehat{\widetilde{\mathrm{MSE}}}}_{\mathrm{HT}}$: the simulation mean for the $\widehat{\widetilde{\mathrm{MSE}}}_{\mathrm{HT}}$ estimator of $\widetilde{\mathrm{MSE}}_{\mathrm{HT}}$ where

$$\overline{\widehat{\widetilde{\mathrm{MSE}}}}_{\mathrm{HT}} = \frac{1}{10{,}000} \sum_{s=1}^{10{,}000} \left(\hat{V}_{\mathrm{Approx},\, s} + \sum_{k \in s} d_k \hat{\sigma}_k^2 \left(d_k - 1\right) \right).$$

- $\overline{\widehat{\mathrm{Weff}}}$: the simulation mean for the $\widehat{\mathrm{Weff}}_S$ estimator (3.4) of Weff.
- $\mathrm{MSE}\left(\widehat{\mathrm{Weff}}_S\right)$: the MSE of $\widehat{\mathrm{Weff}}_S$ simulations defined by

$$\mathrm{MSE}\left(\widehat{\mathrm{Weff}}_S\right) = \frac{1}{10{,}000} \sum_{s=1}^{10{,}000} \left(\widehat{\mathrm{Weff}}_s - \mathrm{Weff}\right)^2.$$

The simulation results for heteroskedastic regression models are presented in Table 4.1 below, while the results for homoskedastic models are given in Table A.1 in the appendix.

Table 4.1
(*Heteroskedastic populations*): Simulation results for the $\widehat{\text{Weff}}$ criterion, by sample size and degree of the link between the variables of interest and the calibration variables

		Variables of interest					
		Y_1	Y_2	Y_3	Y_4	Y_5	Y_6
		($R^2 = 0.01$)	($R^2 = 0.10$)	($R^2 = 0.20$)	($R^2 = 0.50$)	($R^2 = 0.75$)	($R^2 = 0.98$)
$n = 100$	MSE_{Cal} (10^7)	12,301.13	9,334.81	1,860.23	173.61	59.47	3.07
	MSE_{HT} (10^7)	11,285.46	8,643.37	1,841.84	323.46	212.69	160.35
	$\widetilde{\text{MSE}}_{\text{HT}}$ (10^7)	11,285.44	8,643.34	1,841.81	323.43	212.66	160.32
	Weff	1.09	1.08	1.01	0.54	0.28	0.02
	$\overline{\widehat{\text{MSE}}}_{\text{Cal}}$ (10^7)	12,463.22	9,484.87	1,984.51	180.37	62.07	3.21
	$\overline{\widehat{\text{MSE}}}_{\text{HT}}$ (10^7)	11,856.45	9,068.99	1,929.87	330.59	215.13	160.07
	$\overline{\widehat{\text{Weff}}}$	1.08	1.07	1.00	0.55	0.30	0.02
	$\text{MSE}\left(\widehat{\text{Weff}}\right)$	0.030	0.034	0.030	0.02	0.008	0.00005
$n = 200$	MSE_{Cal} (10^7)	5,931.78	4,500.60	905.42	81.86	27.99	1.41
	MSE_{HT} (10^7)	5,543.74	4,245.87	904.76	158.89	104.48	78.77
	$\widetilde{\text{MSE}}_{\text{HT}}$ (10^7)	5,543.72	4,245.85	904.75	158.88	104.46	78.75
	Weff	1.07	1.06	1.00	0.52	0.27	0.02
	$\overline{\widehat{\text{MSE}}}_{\text{Cal}}$ (10^7)	5,770.29	4,382.31	969.57	83.81	28.68	1.48
	$\overline{\widehat{\text{MSE}}}_{\text{HT}}$ (10^7)	5,673.08	4,341.19	924.64	160.71	105.06	78.71
	$\overline{\widehat{\text{Weff}}}$	1.05	1.05	1.01	0.53	0.28	0.02
	$\text{MSE}\left(\widehat{\text{Weff}}\right)$	0.008	0.008	0.007	0.006	0.002	0.00005
$n = 400$	MSE_{Cal} (10^7)	3,847.61	2,919.12	589.97	53.05	18.13	0.94
	MSE_{HT} (10^7)	3,629.83	2,780.03	592.40	104.04	68.41	51.57
	$\widetilde{\text{MSE}}_{\text{HT}}$ (10^7)	3,629.82	2,780.02	592.39	104.03	68.40	51.56
	Weff	1.06	1.05	0.99	0.51	0.27	0.02
	$\overline{\widehat{\text{MSE}}}_{\text{Cal}}$ (10^7)	3,718.79	2,889.81	594.01	53.89	18.44	0.95
	$\overline{\widehat{\text{MSE}}}_{\text{HT}}$ (10^7)	3,687.44	2,821.34	602.39	104.83	68.68	51.60
	$\overline{\widehat{\text{Weff}}}$	1.04	1.04	0.98	0.52	0.27	0.02
	$\text{MSE}\left(\widehat{\text{Weff}}\right)$	0.004	0.005	0.004	0.003	0.001	0.00001

Hence, the simulation results show that the Weff criterion proposed to measure the impact of using calibration weights helps us to identify situations where calibration weighting should not be used, i.e., when the variable of interest is weakly correlated with the calibration variables $(R^2 < 0.20)$. Furthermore, the $\widehat{\text{Weff}}_S$ estimator (3.4) proposed to estimate the Weff criterion proved to be an effective estimator, recording the same performances, regardless of the strength of the link between the variable of interest and the calibration variables. Heteroskedastic residuals for regression models, representing the link between the variable of interest and the calibration variables, had little impact on the performances of the Weff criterion and the $\widehat{\text{Weff}}_S$ estimator. We also noted a lack of impact in using approximation (2.8) for the variance under design $\sum_{k \in S} d_k \mathbf{x}'_k \boldsymbol{\beta}$ since the impact of the deviation between the AMSE for the HT estimator (MSE_{HT}) and its approximation $\widetilde{\text{MSE}}_{\text{HT}}$ (2.10) was negligible in the results for the Weff criterion. This was predictable since the design being considered was a SRSD.

5 Conclusion

In this paper, we have proposed a new criterion for measuring the impact of using calibration weights to estimate the total for a variable of interest. This criterion can be calculated for each variable of interest to determine whether it is better to use a set of calibration weights or sampling weights to estimate the total for the variable. The proposed criterion has the benefit of taking into account the two main aspects that influence the precision of a total estimator: bias due to the use of calibration weights and the quality of the linear regression model that represents the link between the variable of interest and the calibration variables. Therefore, this criterion can be seen as a measurement of the threshold where the gain in the variance obtained with the calibration estimator exceeds the loss in bias due to the use of calibration weights rather than sampling weights. The simulations conducted to evaluate the proposed criterion showed that this criterion does indeed identify, for a given variable of interest, situations where it is best to use calibration weights, i.e., when the variable of interest is sufficiently correlated with the calibration variables.

It is important to note that the role of this criterion is not to introduce a new weighting system to replace calibration weighting or sample weighting. It is used solely to identify which of the two weighting systems would be best to use for a given variable of interest, which is very useful for practitioners, particularly in the case of surveys that cover different subjects, such as omnibus surveys. However, it would be interesting to study the possibility of producing a unique new weighting system for all survey variables, based on this criterion, while taking into account the advantages of both calibration weights and sampling weights. Finally, it should be noted that the proposed criterion requires a linear relationship between the variables of interest and the calibration variables, and the robustness of the criterion is worth investigating.

Acknowledgements

We would like to thank the reviewers for their thorough work, which helped to improve the results presented in this paper.

Appendix

Simulations results for homoskedastic residual models

Table A.1
(*Homoskedastic populations*): Simulation results for the $\widehat{\text{Weff}}$ criterion, by sample size and degree of the link between the variables of interest and the calibration variables

		Variables of interest					
		Y_1 ($R^2 = 0.01$)	Y_2 ($R^2 = 0.10$)	Y_3 ($R^2 = 0.20$)	Y_4 ($R^2 = 0.50$)	Y_5 ($R^2 = 0.75$)	Y_6 ($R^2 = 0.98$)
$n = 100$	MSE_{Cal} (10^7)	30,150.81	9,298.14	1,492.16	177.42	56.54	3.58
	MSE_{HT} (10^7)	27,162.87	8,530.43	1,477.41	326.93	207.72	160.37
	$\widetilde{\text{MSE}}_{\text{HT}}$ (10^7)	27,162.82	8,530.40	1,477.39	326.90	207.69	160.34
	Weff	1.11	1.09	1.01	0.54	0.27	0.02
	$\overline{\widehat{\text{MSE}}}_{\text{Cal}}$ (10^7)	31,523.63	9,775.29	1,565.31	192.17	61.49	3.90
	$\overline{\widehat{\text{MSE}}}_{\text{HT}}$ (10^7)	29,024.17	9,128.96	1,573.25	338.45	211.87	160.75
	$\overline{\widehat{\text{Weff}}}$	1.09	1.07	1.00	0.58	0.30	0.02
	$\text{MSE}\left(\widehat{\text{Weff}}\right)$	0.020	0.021	0.021	0.016	0.007	0.00008
$n = 200$	MSE_{Cal} (10^7)	14,277.16	4,441.79	732.99	83.44	26.59	1.68
	MSE_{HT} (10^7)	13,343.16	4,190.39	725.75	160.60	102.04	78.78
	$\widetilde{\text{MSE}}_{\text{HT}}$ (10^7)	13,343.14	4,190.37	725.73	160.58	102.02	78.77
	Weff	1.07	1.06	1.01	0.52	0.26	0.02
	$\overline{\widehat{\text{MSE}}}_{\text{Cal}}$ (10^7)	14,195.90	4,398.60	753.49	86.72	27.69	1.75
	$\overline{\widehat{\text{MSE}}}_{\text{HT}}$ (10^7)	13,795.17	4,336.28	748.77	163.53	102.90	78.84
	$\overline{\widehat{\text{Weff}}}$	1.06	1.05	1.01	0.53	0.27	0.02
	$\text{MSE}\left(\widehat{\text{Weff}}\right)$	0.003	0.003	0.004	0.005	0.002	0.00002
$n = 400$	MSE_{Cal} (10^7)	9,086.04	2,826.00	470.43	53.96	17.20	1.09
	MSE_{HT} (10^7)	8,736.60	2,743.71	475.19	105.15	66.81	51.58
	$\widetilde{\text{MSE}}_{\text{HT}}$ (10^7)	8,736.58	2,743.69	475.18	105.14	66.80	51.57
	Weff	1.04	1.03	0.99	0.51	0.26	0.02
	$\overline{\widehat{\text{MSE}}}_{\text{Cal}}$ (10^7)	9,178.88	2,894.26	478.67	55.38	17.65	1.12
	$\overline{\widehat{\text{MSE}}}_{\text{HT}}$ (10^7)	8,946.42	2,833.29	485.09	106.41	67.21	51.57
	$\overline{\widehat{\text{Weff}}}$	1.03	1.02	0.98	0.52	0.27	0.02
	$\text{MSE}\left(\widehat{\text{Weff}}\right)$	0.001	0.001	0.002	0.003	0.002	0.00001

References

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

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

Hájek, J. (1981). *Sampling from a Finite Population*. New York: Marcel Dekker.

Henry, K.A., and Valliant, R. (2015). A design effect measure for calibration weighting in single-stage samples. *Survey Methodology*, 41, 2, 315-331. Paper available at https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2015002/article/14236-eng.pdf.

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

Matei, A., and Tillé, Y. (2005). Evaluation of variance approximations and estimators in maximum entropy sampling with unequal probability and fixed sample size. *Journal of Official Statistics*, 21(4), 2005, 543-570.

Nedyalkova, D., and Tillé, Y. (2008). Optimal sampling and estimation strategies under linear model. *Biometrika*, 95, 521-537.

Tirari, M.H.T. (2003). Estimation d’un total pour les plans de sondage à taille fixe et équilibrés. Thesis report.

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following people who have provided help or served as referees for one or more papers during 2018.

S. Amine, *Université du Québec en Outaouais*
P. Biemer, *RTI*
C. Bocci, *Statistics Canada*
C. Boulet, *Statistics Canada*
F. Brisebois, *Statistics Canada*
G. Chauvet, *ENSAI/IRMAR*
J. Chiche, *CEVIPOF*
T. Deroyon, *INSEE*
L. Dudoignon, *Médiamétre*
M. El Haj Tirari, *INSEE*
S. Falorsi, *ISTAT*
C. Goga, *Université de Franche-Comté*
A. Goia, *Dipartimento di studi per l'economia et l'impresa*
A. Grafström, *Swedish university of agricultural sciences*
E. Gros, *INSEE*
S. Hallépée, *INSEE*
J. Im, *Iowa State University*
J. Legg, *Amgen Inc.*
S. Legleye, *INSEE*
É. Lesage, *Statistics Canada*
R. Lethonen, *University of Helsinki*
A.G. Matei, *Université de Neuchâtel*
J. Opsomer, *Westat*
G. Santin, *Agence de la biomédecine*
Y. Tillé, *Université de Neuchâtel*
J. Van den Brakel, *Statistics Netherland and Maastricht University*
Y. You, *Statistics Canada*

Acknowledgements are also due to those who assisted during the production of the 2018 issues: Céline Ethier of International Cooperation and Corporate Statistical Methods Division; Joana Bérubé of Business Survey Methods Division; Gilbert Côté of Official Release, Language Services and Respondent Communications Division; the team from Dissemination Division, in particular: Chantal Chalifoux, Christina Jaworski, Kathy Charbonneau, Jacqueline Luffman, Isabelle Gravelle and Joseph Prince as well as our partners in the Communications Division.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 34, No. 2, June 2018

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 34, No. 3, September 2018

All inquires about submissions and subscriptions should be directed to jos@scb.se

GUIDELINES FOR MANUSCRIPTS

Authors are invited to submit their articles in English or French in electronic form to the Editor, (statcan.smj-rte.statcan@canada.ca). Before submitting the article, please examine a recent issue of *Survey Methodology* (Vol. 39, No. 1 and onward) as a guide and note particularly the points below. Articles must be submitted in machine-readable form, preferably in Word with MathType for the mathematical expressions. A pdf or paper copy may be required for formulas and figures.

1. Layout

1.1 Documents should be typed entirely double spaced with margins of at least 1½ inches on all sides.
1.2 The documents should be divided into numbered sections with suitable verbal titles.
1.3 The name (fully spelled out) and address of each author should be given as a footnote on the first page of the manuscript.
1.4 Acknowledgements should appear at the end of the text.
1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

3.1 Avoid footnotes, abbreviations, and acronyms.
3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “exp(·)” and “log(·)”, etc.
3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered with arabic numerals on the right if they are to be referred to later. Use a two-level numbering system based on the section of the paper. For example, equation (4.2) is the second important equation in section 4.
3.4 Write fractions in the text using a solidus.
3.5 Distinguish between ambiguous characters, (e.g., w, ω; o, O, 0; l, 1).
3.6 If possible, avoid using bold characters in formulae.

4. Figures and Tables

4.1 All figures and tables should be numbered with arabic numerals, with titles that are as self explanatory as possible, at the bottom for figures and at the top for tables. Use a two-level numbering system based on the section of the paper. For example, table 3.1 is the first table in section 3.
4.2 A detailed textual description of figures may be required for accessibility purposes if the message conveyed by the image is not sufficiently explained in the text.

5. References

5.1 References in the text should be cited with authors’ names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, page 164).
5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

6. Short Notes

6.1 Documents submitted for the short notes section must have a maximum of 3,000 words.