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

Survey Methodology

Survey Methodology
46-1

Release date: June 30, 2020

Statistics Canada
Statistique Canada

Canadä

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.

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology

Catalogue No. 12-001-XPB

A journal published by Statistics Canada

June 2020

•

Volume 46

•

Number 1

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

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

EDITORIAL POLICY

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

Submission of Manuscripts

Survey Methodology is published twice a year in electronic format. Authors are invited to submit their articles **through the *Survey Methodology* hub on the ScholarOne Manuscripts website** (https://mc04.manuscriptcentral.com/surveymeth). For formatting instructions, please see the guidelines provided in the journal and on the web site (www.statcan.gc.ca/surveymethodology). To communicate with the Editor, please use the following email: (statcan.smj-rte.statcan@canada.ca).

Survey Methodology

A Journal Published by Statistics Canada

Volume 46, Number 1, June 2020

Contents

Regular Papers

Survey Methodology, June 2020
Vol. 46, No. 1, pp. 1-28
Statistics Canada, Catalogue No. 12-001-X

Are probability surveys bound to disappear for the production of official statistics?

Jean-François Beaumont[1]

Abstract

For several decades, national statistical agencies around the world have been using probability surveys as their preferred tool to meet information needs about a population of interest. In the last few years, there has been a wind of change and other data sources are being increasingly explored. Five key factors are behind this trend: the decline in response rates in probability surveys, the high cost of data collection, the increased burden on respondents, the desire for access to "real-time" statistics, and the proliferation of non-probability data sources. Some people have even come to believe that probability surveys could gradually disappear. In this article, we review some approaches that can reduce, or even eliminate, the use of probability surveys, all the while preserving a valid statistical inference framework. All the approaches we consider use data from a non-probability source; data from a probability survey are also used in most cases. Some of these approaches rely on the validity of model assumptions, which contrasts with approaches based on the probability sampling design. These design-based approaches are generally not as efficient; yet, they are not subject to the risk of bias due to model misspecification.

Key Words: Statistical matching; Calibration; Non-probabilistic data; Data integration; Fay-Herriot model; Propensity score.

1 Introduction

In 1934, Jerzy Neyman laid the foundation for probability survey theory and his design-based approach to inference with an article published in the *Journal of the Royal Statistical Society*. His article (Neyman, 1934) piqued the interest of a number of statisticians at the time, and the theory was developed further in the following years. Still today, many articles on this topic are published in statistics journals. Rao (2005) provides an excellent review of various developments in probability survey theory during the 20th century (see also Bethlehem, 2009; Rao and Fuller, 2017; Kalton, 2019). Nowadays, national statistical agencies, such as Statistics Canada and the Institut National de la Statistique et des Études Économiques (INSEE) in France, use probability surveys most often to get the information they seek on a population of interest.

The popularity of probability surveys for producing official statistics stems largely from the non-parametric nature of the inference approach developed by Neyman (1934). In other words, probability surveys allow for valid inferences about a population without having to rely on model assumptions. This is appealing – even fundamental, according to Deville (1991) – to national statistical agencies that produce official statistics. In fact, these agencies have historically been reluctant to take unnecessary risks, which are unavoidable for approaches that depend on the validity of model assumptions, especially when it is difficult to check the underlying assumptions.

[1] Jean-François Beaumont, Statistics Canada, R.H. Coats Building, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6. E-mail: jean-francois.beaumont@canada.ca.

However, estimates from probability surveys can prove inefficient, even to the point of being unusable, particularly when the sample size is small (see, for example, Rao and Molina, 2015). Furthermore, they are based on the assumption that non-sampling errors, such as measurement, coverage or non-response errors, are negligible. To minimize these errors, national statistical agencies often invest considerable resources. For example, questionnaires are tested to ensure that respondents fully understand them; survey data are validated using various edit rules; respondents are contacted again, if necessary, to confirm the data collected; non-respondent follow-ups are conducted to minimize the impact of non-response on the estimates, etc. Despite all these efforts, non-sampling errors persist in practice. There are, of course, adaptations of the theory for taking these errors into account. These adaptations necessarily come with model assumptions and thus with the risk of bias resulting from inadequate assumptions. Probability surveys are not a panacea but they are generally recognized as providing a reliable source of information about a population, except when non-sampling errors become dominant. Brick (2011) takes the argument further and defends the idea that a probability survey with a low response rate – if properly designed – usually provides estimates with smaller bias than those obtained from a volunteer non-probability survey. Dutwin and Buskirk (2017) show empirical results that corroborate this argument.

For the past few years, a wind of change has been blowing over national statistical agencies, and other data sources are being increasingly explored. Five key factors are behind this trend: i) the decline in response rates in probability surveys in recent years; ii) the high cost of data collection; iii) the increased burden on respondents; iv) the desire for access to “real-time” statistics (Rao, 2020), in other words, having the ability to produce statistics practically at the same time or very shortly after the information needs are expressed; and v) the proliferation of non-probability data sources (Rancourt, 2019) such as administrative sources, social media, web surveys, etc. To control data collection costs of probability surveys and reduce the adverse effects of non-response on the quality of estimates, a number of authors have proposed and evaluated responsive data collection methods (e.g., Laflamme and Karaganis, 2010; Lundquist and Särndal, 2013; Schouten, Calinescu and Luiten, 2013; Beaumont, Haziza and Bocci, 2014; Särndal, Lumiste and Traat, 2016). Tourangeau, Brick, Lohr and Li (2017) review various methods and point out their limited success in reducing non-response bias and costs. Särndal et al. (2016) also reach the same conclusion regarding bias. Some surveys conducted by national statistical agencies still have very low response rates, and it becomes risky to rely solely on data collection and estimation methods to correct potential non-response biases. Indeed, a number of authors (e.g., Rivers, 2007; Elliott and Valliant, 2017) pointed out the similarity between a probability survey with a very low response rate and a non-probability survey. Yet, a non-probability survey has the advantages of having a usually much larger sample size and being less costly. Given the above discussion, some have come to believe that probability surveys could gradually disappear (see Couper, 2000; Couper, 2013; Miller, 2017).

However, data from non-probability sources are not without challenges, as noted by Couper (2000), Baker, Brick, Bates, Battaglia, Couper, Dever, Gile and Tourangeau (2013), and Elliott and Valliant (2017), among others. For example, it is well known that non-probability surveys that collect data from volunteers can often lead to estimates with significant selection bias (or participation bias). Bethlehem (2016) provides a bias expression and argues that the potential for bias is usually higher for a non-probability survey than for a probability survey affected by non-response. Meng (2018) illustrates that bias becomes dominant as the non-probability sample size increases, which significantly reduces the effective sample size. Therefore, the acquisition of large non-probability samples alone cannot ensure the production of estimates with an acceptable quality. The pre-election poll conducted by the *Literary Digest* magazine for predicting the outcome of the 1936 U.S. presidential election is a prime example of this (Squire, 1988; Elliott and Valliant, 2017). Despite a huge sample size of over two million people, the poll was unable to predict Franklin Roosevelt's overwhelming victory. Instead, it incorrectly predicted a convincing victory for his opponent, Alfred Landon. The set of poll respondents, who were highly unrepresentative of the voting population, was made up mainly of car and phone owners as well as the magazine's subscribers. Couper (2000) and Elliott and Valliant (2017) cite other more recent examples of non-probability surveys that led to erroneous conclusions.

Selection bias is not the only challenge that must be overcome when using data from a non-probability source. Another major challenge is the presence of measurement errors (e.g., Couper, 2000). They can significantly impact the estimates, especially when data are collected without relying on an experienced interviewer. This is the case for most non-probability sources, in particular volunteer web surveys.

The current context leads to the following question: How can data from a non-probability source be used to minimize, even eliminate, the data collection costs and respondent burden of a probability survey, all the while preserving a valid statistical inference framework and acceptable quality? That is the main question this article attempts to answer.

Most of the methods we present integrate data from a probability survey and a non-probability source. Zhang (2012) discusses the concept of statistical validity when integrated data are used to make inferences. We contend that establishing a statistical framework that can be used to make valid inferences is essential for the production of official statistics, a point that also seems to be shared by Rancourt (2019). Without such a framework, the usual properties of estimators, such as bias and variance, are not defined. It then becomes impossible to select estimators based on an objective criterion such as, for example, choosing the linear unbiased estimator with the smallest possible variance. Without a valid statistical inference framework, estimates can be calculated, but all the usual tools for determining the quality of those estimates and drawing accurate conclusions about the population's characteristics of interest are lost.

In the rest of this article, we differentiate design-based approaches to inference, described in Section 3, from model-based approaches to inference, described in Section 4. For each approach, we consider two scenarios: In the first one, the data from the non-probability source match exactly the concepts of interest and are not fraught with measurement errors. Those data can therefore be used to replace the data from a probability survey. In the second scenario, the data from the non-probability source do not reflect concepts of interest or are subject to measurement errors. Although these data cannot be used to directly replace data from a probability survey, they can still be used as auxiliary information to enhance it. In Section 5, we provide some additional thoughts. Let us first begin with some background in Section 2.

2 Background

One of the first steps to meet information needs is to define the target population for which that information is sought. We denote this target population by U. Then, it is necessary to define the parameters of interest, i.e. what it is desired to know about the target population. In practice, it is often desired to estimate many parameters. To simplify the discussion, we suppose that only one parameter is of interest: the total of the variable y, $\theta = \sum_{k \in U} y_k$, where y_k is the value of the variable y for unit k of the population U. We use $\mathbf{Y}$ to denote the vector containing the values y_k for $k \in U$. Lastly, a set of procedures must be established for the estimation of the parameter θ while taking into account various factors, such as the available budget, the respondent burden, the desired precision, etc. During this process, it is necessary to identify the data sources that will be used – probabilistic or not – and a statistical inference framework that will allow for assessing the properties of the estimates produced, such as bias and variance.

The above sequence, which starts with defining the target population and parameters of interest, followed by the data sources and estimation procedures, is consistent with the proposal by Citro (2014). She suggests that national statistical agencies first determine the information needs along with potential users. Next, they can work at identifying the data source(s) that will meet those needs while preserving an acceptable quality of estimates, keeping costs within the established budget and controlling for respondent burden. It seems preferable to avoid the reverse procedure, however tempting it is, of first identifying available data sources and then artificially determining the needs based on what can be produced by these sources. In general, this kind of procedure cannot adequately meet users' actual needs.

We assume that we have access to data from a non-probability source (e.g., administrative data, web survey data, etc.). Values are observed for a few variables, including a variable y^*, for all units of a subset of U, denoted as s_{NP}. The variable y^* is not necessarily equal to y because of conceptual differences and/or measurement errors. At least, it is hoped that there is a strong association between the two variables. We denote the inclusion indicator in s_{NP} as δ_k; in other words, $\delta_k = 1$ if unit k is in s_{NP} and $\delta_k = 0$, otherwise. The vector of the inclusion indicators δ_k for $k \in U$ is denoted by $\boldsymbol{\delta}$.

Data from a probability survey may also be available. In that case, a sample s_P of the population U is randomly selected with probability $p(s_P \mid \mathbf{Z})$. The matrix $\mathbf{Z}$ contains information available on the sampling frame that is used to define the sampling design, such as stratum identifiers for each unit of the population. The sample inclusion indicators, I_k, $k \in U$, are defined as follows: $I_k = 1$ if unit k is selected in the sample s_P; otherwise, $I_k = 0$. We use $\mathbf{I}$ to denote the vector containing the sample inclusion indicators for $k \in U$. The probability that unit k of the population U is chosen in the sample is denoted by $\pi_k = E(I_k \mid \mathbf{Z})$. Most of the time it is known or can be approximated. We assume that

$\pi_k > 0, k \in U$. For each unit $k \in s_P$, the values of certain variables are collected, which may or may not include the variable y.

We use $\mathbf{\Omega}$ to denote the set of all the auxiliary data used to make inferences. Among other things, $\mathbf{\Omega}$ includes the design information, $\mathbf{Z}$, if a probability sample is used, and potentially other auxiliary variables such as calibration variables, matching variables or explanatory variables of a model (see Sections 3 and 4). The inclusion indicator δ_k can also be used as an auxiliary variable either for stratifying the population or for calibration (see Section 3). The vector $\boldsymbol{\delta}$ can thus be included in $\mathbf{Z}$ and $\mathbf{\Omega}$.

The following two assumptions are used throughout the article:

Assumption 1: $\mathbf{I}$ is independent of $\mathbf{\Omega}$ and $\mathbf{Y}$ after conditioning on $\mathbf{Z}$.
Assumption 2: $\boldsymbol{\delta}$ and $\mathbf{I}$ are independent after conditioning on $\mathbf{\Omega}$ and $\mathbf{Y}$.

Assumption 1 implies that the values of the variables included in $\mathbf{\Omega}$ and $\mathbf{Y}$ are not affected by whether or not a unit is included in the sample s_P. This is implicit in the literature on probability surveys and results from the very definition of the sampling design, which depends only on $\mathbf{Z}$. Assumption 2 is automatically satisfied if the non-probability source (and thus $\boldsymbol{\delta}$) is available prior to selecting the probability sample. Note that if δ_k is used as an auxiliary variable to stratify the population, then $\boldsymbol{\delta}$ is included in $\mathbf{\Omega}$ and assumption 2 is still satisfied. It will not be satisfied if being selected in s_P impacts the provision of data to the non-probability source. For example, being selected in s_P (and contacted) could be an indirect reminder for the selected individual to fill out forms required by the government (non-probability source). It can be expected that assumptions 1 and 2 are satisfied in most cases.

The union of $\mathbf{\Omega}, \boldsymbol{\delta}, \mathbf{I}$ and $\mathbf{Y}$ contains all the information used for making inferences. The various approaches to inference set out in Sections 3 and 4 differ in what they treat as fixed and what they treat as random. For example, in the design-based approach to inference, everything is considered fixed except for the vector $\mathbf{I}$; in other words, design-based inferences are conditional on $\mathbf{\Omega}, \mathbf{Y}$ and $\boldsymbol{\delta}$. To simplify the notation, we use $\mathbf{\Omega}_P$ to denote the union of $\mathbf{\Omega}, \mathbf{Y}$ and $\boldsymbol{\delta}$. Thus, design expectations are denoted as $E(\cdot \mid \mathbf{\Omega}_P)$ rather than $E(\cdot \mid \mathbf{\Omega}, \mathbf{Y}, \boldsymbol{\delta})$. In the design-based approach to inference, an estimator $\hat{\theta}$ of θ is usually chosen so that the design bias, $E\left(\hat{\theta} - \theta \,\middle|\, \mathbf{\Omega}_P\right)$, is zero or negligible. Under assumptions 1 and 2, we note that $E(I_k \mid \mathbf{\Omega}_P) = E(I_k \mid \mathbf{Z}) = \pi_k$. For estimating the total $\theta = \sum_{k \in U} y_k$, an estimator of the form $\hat{\theta} = \sum_{k \in s_P} w_k y_k$ is frequently used, where w_k is a survey weight for unit k. The standard basic weight is $w_k = \pi_k^{-1}$. This weight ensures that the estimator $\hat{\theta}$ is exactly design-unbiased for θ. The basic weight can then be modified using calibration techniques (e.g., Deville and Särndal, 1992; Haziza and Beaumont, 2017). The advantage of this approach is its non-parametric nature: no model assumption is needed for making valid inferences about the population because the first two design moments are controlled by the statistician and are usually known. Yet, the approach is not free of assumptions, for example to ensure the consistency and asymptotic normality of estimators, but it does not require any parametric model.

In practice, non-response is often observed in probability surveys as well as other non-sampling errors. Non-response of some sample units is often viewed as an additional phase of sampling that is not controlled by the statistician. In other words, the non-response mechanism is not known, unlike the sampling design. Assuming an adequate model for the non-response mechanism, estimators with little or no bias can be obtained, for example by weighting the responding units by the inverse of their estimated response probability. However, this requires careful modelling of the response indicators. In the rest of this paper, we ignore non-sampling errors and assume that the estimates from the probability survey are not biased or, at least, that their bias is small compared to the bias of the estimates from the non-probability source alone. This assumption may not always be satisfied in practice, but it is reasonable in many contexts (see Brick, 2011), especially in large surveys conducted by national statistical agencies.

The acquisition of data from non-probability sources is generally inexpensive compared to the cost of collecting data from a probability survey. Therefore, they would ideally be used to replace data from a probability survey. This data replacement is valid only if $y_k^* \approx y_k, k \in U$. This assumption will not be satisfied with all non-probability data sources, but may be realistic with some administrative data sources. In Sections 3 and 4, we will differentiate the methods based on the assumption that $y_k^* = y_k$ from the methods not requiring this assumption. Several methods described in Sections 3 and 4 are also reviewed in the upcoming article by Rao (2020) that was presented to Statistics Canada in summer 2018.

3 Design-based approaches

Design-based approaches yield design-consistent estimators of θ even when the non-probability source produces estimates with a significant selection bias. In this context, the purpose of using a non-probability sample is to reduce the variance of estimators of θ. The efficiency gains achieved can be used to justify a reduction of the probability sample size, thereby a reduction of the data collection costs and respondent burden. The methods that we consider in Sections 3.1 and 3.2 require collecting the values of the variable of interest y in the probability sample, just like small area estimation methods described in Section 4.4. However, the efficiency gains are usually expected to be more modest than those obtained using small area estimation methods. In Section 3.1, we consider the scenario $y_k^* = y_k$ whereas in Section 3.2, we consider the scenario $y_k^* \neq y_k$.

3.1 Weighting by the inverse of the probability of inclusion in the combined sample

The ideal case occurs when the non-probability sample is a census, i.e., $s_{\mathrm{NP}} = U$. In that case, the value of the parameter of interest $\theta = \sum_{k \in U} y_k$ can be directly calculated without worrying about bias or variance since $y_k^* = y_k$ is assumed in this section. In general, we expect under-coverage in the sense that s_{NP} is smaller than the population U. In a design-based approach, the potential under-coverage bias can be addressed by selecting a probability sample s_P from U and collecting the values of the variable y for the sample units. Ideally, the probability sample is drawn from $U - s_{\mathrm{NP}}$ but it is possible that the units in s_{NP} cannot be linked to those of the sampling frame U to establish the set $U - s_{\text{NP}}$. In general, the larger the non-probability sample, the more it is possible to reduce the size of the probability sample without jeopardizing the desired precision of the estimates.

It seems desirable to estimate θ using all the data collected in the combined sample $s = s_P \cup s_{\text{NP}}$. The inclusion indicator in s can be defined as $\tilde{I}_k = \delta_k + (1 - \delta_k) I_k$. To obtain a design-unbiased estimator of θ, each unit $k \in s$ is weighted by $\tilde{w}_k = \tilde{\pi}_k^{-1}$ where $\tilde{\pi}_k = E(\tilde{I}_k \mid \mathbf{\Omega}_P)$. Under assumptions 1 and 2, $E(I_k \mid \mathbf{\Omega}_P) = \pi_k$ and we obtain

$$\tilde{\pi}_k = E(\tilde{I}_k \mid \mathbf{\Omega}_P) = \delta_k + (1 - \delta_k) \pi_k.$$

The resulting estimator is written:

$$\hat{\theta} = \sum\nolimits_{k \in s} \tilde{w}_k y_k = \sum\nolimits_{k \in s_{\text{NP}}} y_k + \sum\nolimits_{k \in s_P} \frac{1}{\pi_k} (1 - \delta_k) y_k. \tag{3.1}$$

Note that estimator (3.1) requires the indicator δ_k to be available for all units in the sample s_P. For the units $k \in s_P \cap s_{\text{NP}}$, we have two values: y_k and y_k^*. In principle, we should have $y_k^* = y_k$, but it is possible that this relationship is not exactly satisfied. These units can be used to validate the assumption $y_k^* \approx y_k$. If significant differences are observed, it may be preferable to not consider this approach and to rely on the methods in Section 3.2 that use data from the non-probability source as auxiliary data. If we trust the data quality of the non-probability source, it may be advisable not to collect the variable y in the probability sample for the units also present in the non-probability sample in order to reduce the data collection costs and respondent burden.

We can view the problem as if we had two sampling frames: U and s_{NP}. A sample s_P is drawn randomly from U and a census is taken from s_{NP}. The probability of selection in the sample s, $\Pr(k \in s \mid \mathbf{\Omega}_P)$, can then be calculated for each unit $k \in U$, and the estimator (3.1) is recovered by weighting each unit $k \in s$ by the inverse of that probability. This approach was proposed by Bankier (1986) to address the problem of multiple sampling frames. In the context of integrating a probability and non-probability sample, estimator (3.1) was proposed by Kim and Tam (2020).

The last sum of (3.1) is a design-unbiased estimator of $\sum\nolimits_{k \in U} (1 - \delta_k) y_k = \sum\nolimits_{k \in U - s_{\text{NP}}} y_k$. If a vector of auxiliary variables, $\mathbf{x}_k$, is available for $k \in s_P$ as well as the total $\mathbf{T}_{\mathbf{x}} = \sum\nolimits_{k \in U} \mathbf{x}_k$ then the weight $1/\pi_k$ in (3.1) can be replaced with a calibrated weight w_k (e.g., Deville and Särndal, 1992; Haziza and Beaumont, 2017). The calibrated weights minimize a distance function between w_k and $1/\pi_k$, $k \in s_P$, under the constraint of satisfying the calibration equation $\sum\nolimits_{k \in s_P} w_k \mathbf{x}_k = \mathbf{T}_{\mathbf{x}}$. Ideally, the calibration is done only on the portion not covered by the non-probability sample, $U - s_{\text{NP}}$; i.e., the calibration vector $(1 - \delta_k) \mathbf{x}_k$ is used, and the calibration equation becomes: $\sum\nolimits_{k \in s_P} w_k (1 - \delta_k) \mathbf{x}_k = \sum\nolimits_{k \in U - s_{\text{NP}}} \mathbf{x}_k$. This is not possible when $\sum\nolimits_{k \in U - s_{\text{NP}}} \mathbf{x}_k$ is unknown.

Remark: If assumption 2 is not appropriate, then $E(I_k \mid \mathbf{\Omega}_P) \neq E(I_k \mid \mathbf{Z}) = \pi_k$. To get around this problem, all the units for which the data were collected after selecting the sample s_P can be removed from s_{NP}. Assumption 2 is then satisfied, but a lot of available data may be omitted. To take advantage of the full set s_{NP}, it is necessary to make a few assumptions and partially depart from the design-based approach. Assuming that $E(I_k \mid \mathbf{\Omega}_P) = \Pr(I_k = 1 \mid \delta_k, \mathbf{Y}, \mathbf{\Omega})$, we can use Bayes' theorem to show that

$$\Pr(I_k = 1 \mid \delta_k = 0, \mathbf{Y}, \mathbf{\Omega}) = \frac{1 - \Pr(\delta_k = 1 \mid I_k = 1, \mathbf{Y}, \mathbf{\Omega})}{1 - \Pr(\delta_k = 1 \mid \mathbf{Y}, \mathbf{\Omega})} \pi_k,$$

for the units $k \in U - s_{\text{NP}}$. Therefore, estimating $E(I_k \mid \mathbf{\Omega}_P)$ requires postulating a model for δ_k. Under some assumptions, $\Pr(\delta_k = 1 \mid I_k = 1, \mathbf{Y}, \mathbf{\Omega})$ can be estimated using the data from the probability sample and, for example, a logistic regression model. Estimating $\Pr(\delta_k = 1 \mid \mathbf{Y}, \mathbf{\Omega})$ can be done using the methods described in Section 4.3 that do not rely on the validity of assumption 2, such as the method by Chen, Li and Wu (2019). These methods require that the auxiliary variables used to model this probability be available for all units of the combined sample $s = s_P \cup s_{\text{NP}}$. Unlike in Section 4.3, here we can take advantage of the availability of y_k for all units of both samples, and we can use the variable of interest as an auxiliary variable. Then, θ is estimated by replacing π_k in (3.1) with an estimate of $\Pr(I_k = 1 \mid \delta_k = 0, \mathbf{Y}, \mathbf{\Omega})$. Similar approaches were proposed by Beaumont, Bocci and Hidiroglou (2014) to take into account late respondents in Statistics Canada's National Household Survey, i.e., households that responded to the initial questionnaire after the follow-up probability sample of non-respondents was drawn.

3.2 Calibration of the probability sample to the non-probability source

Data from non-probability sources, such as those provided by web panel respondents, can be fraught with measurement errors large enough to cast doubt on the assumption that $y_k^* \approx y_k$. Therefore, such data cannot be used to directly replace the values of the variable y. However, they can be used as auxiliary data to enhance the probability survey using the calibration technique. The non-probability source contains the values y_k^* for $k \in s_{\text{NP}}$ and potentially the values of other variables. From all these variables, it is possible to form a vector of auxiliary variables $\mathbf{x}_k^*$, available for $k \in s_{\text{NP}}$, that could include an intercept. Its total is denoted as $\mathbf{T}_{\mathbf{x}^*} = \sum_{k \in s_{\text{NP}}} \mathbf{x}_k^* = \sum_{k \in U} \delta_k \mathbf{x}_k^*$. Another vector of auxiliary variables, $\mathbf{x}_k$, may also be available for $k \in s_P$, as well as its total for the entire population U, $\mathbf{T}_{\mathbf{x}} = \sum_{k \in U} \mathbf{x}_k$. The calibrated weights w_k, $k \in s_P$, are obtained by minimizing a distance function between w_k and $1/\pi_k$, $k \in s_P$, under the constraint of satisfying the calibration equation

$$\sum_{k \in s_P} w_k \begin{pmatrix} \mathbf{x}_k \\ \delta_k \mathbf{x}_k^* \end{pmatrix} = \begin{pmatrix} \mathbf{T}_{\mathbf{x}} \\ \mathbf{T}_{\mathbf{x}^*} \end{pmatrix}.$$

Note that this calibration can be done only if $\mathbf{x}_k^*$ is available in the probability sample for all units $k \in s_P \cap s_{\text{NP}}$. The estimator of θ is again written as $\hat{\theta} = \sum_{k \in s_P} w_k y_k$, where w_k is the calibrated weight satisfying the above calibration equation. No model assumption is required for the validity of the approach, and the resulting estimates remain design-consistent regardless of the strength of the relationship between y_k and the auxiliary variables $\mathbf{x}_k$ and $\mathbf{x}_k^*$. A strong relationship will help reduce the design variance of $\hat{\theta}$, $\operatorname{var}(\hat{\theta} \mid \mathbf{\Omega}_P)$. Kim and Tam (2020) discuss the use of such calibration.

Canada's Labour Force Survey (LFS) provides an example of a potential application for this calibration method. The unemployment rate, defined as the number of unemployed persons divided by the number of persons in the labour force, is a key parameter of interest that the LFS estimates. To improve the precision of the LFS estimates, a calibration variable indicating whether an individual is receiving employment insurance could be effective because there is definitely a connection between receiving employment insurance and being unemployed. The total of this calibration variable, the number of employment insurance beneficiaries, is needed for implementing this calibration and is available from an administrative source. However, applying this method would require adding a question to the LFS to identify LFS respondents who are receiving employment insurance. This information could also be obtained through a linkage between the LFS and the administrative source. It remains to be determined whether such a calibration variable could yield significant gains in the LFS.

4 Model-based approaches

Model-based approaches can eliminate the selection bias of the non-probability source and enable valid statistical inferences, provided that their underlying assumptions hold. The objective of the methods in Sections 4.1, 4.2 and 4.3 is to reduce respondent burden and costs by eliminating data collection for some variables of interest in a probability sample. The greater the number of variables of interest for which the values are not collected, the greater the reduction in data collection costs and respondent burden. However, these methods assume that the variables of interest are measured without error in the non-probability sample $\left(y_k^* = y_k\right)$.

From the non-probability sample s_{NP}, we can obtain the naive estimator $\hat{\theta}^{\text{NP}} = N\sum_{k \in s_{\text{NP}}} y_k / n^{\text{NP}}$ of the total θ, where n^{NP} is the number of units in s_{NP} and N is the size of the population U. It is well known that the selection bias of the naive estimator may be significant (see, for example, Bethlehem, 2016). The objective of the methods in Sections 4.1, 4.2 and 4.3 is to reduce the bias of the naïve estimator by using a vector of auxiliary variables, $\mathbf{x}_k$. We use $\mathbf{X}$ to denote the matrix that contains the values of vector $\mathbf{x}_k$, $k \in U$. We assume that $\mathbf{x}_k$ is measured without error in both samples s_{NP} and s_P.

Section 4.4 briefly discusses small area estimation and the area-level model of Fay and Herriot (1979). Small area estimation methods are generally used to improve the precision of estimates for population sub-groups (domains) that have a small probability sample size. They require collecting the variable y in the probability sample, but not in the non-probability sample. Therefore, they do not require the condition $y_k^* = y_k$. Ideally, the non-probability sample contains variables correlated to y.

4.1 Calibration of the non-probability sample

The most natural approach to correcting the selection bias of a non-probability source is to model the relationship between the variable of interest y_k and the auxiliary variables $\mathbf{x}_k$ and then predict the total θ by predicting the variable y_k for each unit outside the non-probability sample. This prediction approach is described in Royall (1970) and generalized in Royall (1976); see also Elliott and Valliant

(2017). Readers are referred to Valliant, Dorfman and Royall (2000) for more details. With this approach, inferences are conditional on $\boldsymbol{\delta}$ and $\mathbf{X}$. As a result, $\mathbf{Y}$ is considered random as well as $\boldsymbol{\Omega}$ (unless $\boldsymbol{\Omega} = \mathbf{X}$). If a probability sample is used, $\mathbf{I}$ is also considered random. It is usually assumed that the nonprobability sample selection mechanism is not informative:

Assumption 3: $\mathbf{Y}$ and $\boldsymbol{\delta}$ are independent after conditioning on $\mathbf{X}$.

Assumption 3 is the key to eliminating the selection bias. The more access we have to auxiliary variables that are strongly related to both y_k and δ_k, the more plausible assumption 3 becomes. In other words, the richer $\mathbf{X}$ is, the more the conditional independence between $\mathbf{Y}$ and $\boldsymbol{\delta}$ becomes a realistic assumption. This assumption, called the *exchangeability* assumption, is discussed in Mercer, Kreuter, Keeter and Stuart (2017). Schonlau and Couper (2017) also discuss the selection of auxiliary variables and emphasize their key role in reducing selection bias.

Often, a linear model is considered where it is assumed that the observations y_k are mutually independent with $E(y_k \mid \mathbf{X}) = \mathbf{x}_k'\boldsymbol{\beta}$ and $\operatorname{var}(y_k \mid \mathbf{X}) \propto v_k$, where $\boldsymbol{\beta}$ is a vector of unknown model parameters and v_k is a known function of $\mathbf{x}_k$. The best linear unbiased predictor of θ (see, for example, Valliant, Dorfman and Royall, 2000) is given by

$$\hat{\theta}^{\text{BLUP}} = \sum_{k \in s_{\text{NP}}} y_k + \sum_{k \in U - s_{\text{NP}}} \mathbf{x}_k'\hat{\boldsymbol{\beta}} = \mathbf{T}_{\mathbf{x}}'\hat{\boldsymbol{\beta}} + \sum_{k \in s_{\text{NP}}} \left(y_k - \mathbf{x}_k'\hat{\boldsymbol{\beta}}\right), \tag{4.1}$$

where

$$\hat{\boldsymbol{\beta}} = \left(\sum_{k \in s_{\text{NP}}} v_k^{-1}\mathbf{x}_k\mathbf{x}_k'\right)^{-1} \sum_{k \in s_{\text{NP}}} v_k^{-1}\mathbf{x}_k y_k.$$

The predictor $\hat{\theta}^{\text{BLUP}}$ can also be re-written in the weighted form $\hat{\theta}^{\text{BLUP}} = \sum_{k \in s_{\text{NP}}} w_k^C y_k$, where

$$w_k^C = 1 + v_k^{-1}\mathbf{x}_k'\left(\sum_{k \in s_{\text{NP}}} v_k^{-1}\mathbf{x}_k\mathbf{x}_k'\right)^{-1}\left(\mathbf{T}_{\mathbf{x}} - \sum_{k \in s_{\text{NP}}} \mathbf{x}_k\right). \tag{4.2}$$

It can easily be shown that w_k^C is a calibrated weight that satisfies the calibration equation $\sum_{k \in s_{\text{NP}}} w_k^C \mathbf{x}_k = \mathbf{T}_{\mathbf{x}}$. Therefore, the prediction approach is equivalent to calibration when a linear model is used to describe the relationship between y_k and $\mathbf{x}_k$. The calibration equation satisfies what Mercer et al. (2017) call the *composition* assumption. This approach requires knowing the vector of control totals $\mathbf{T}_{\mathbf{x}}$. If it is unknown, an alternative is to replace it in (4.1) or (4.2) with an estimate, $\hat{\mathbf{T}}_{\mathbf{x}} = \sum_{k \in s_P} w_k \mathbf{x}_k$, from a probability survey (Elliott and Valliant, 2017). If assumptions 1 to 3 are satisfied, it can be shown that the predictor $\hat{\theta}^{\text{BLUP}}$ is unbiased, i.e., $E\left(\hat{\theta}^{\text{BLUP}} - \theta \,\middle|\, \boldsymbol{\delta}, \mathbf{X}\right) = 0$, whether $\mathbf{T}_{\mathbf{x}}$ or $\hat{\mathbf{T}}_{\mathbf{x}}$ is used, provided that the latter is design-unbiased, i.e., $E\left(\hat{\mathbf{T}}_{\mathbf{x}} \,\middle|\, \boldsymbol{\Omega}_P\right) = \mathbf{T}_{\mathbf{x}}$. Of course, the unbiasedness property of the predictor $\hat{\theta}^{\text{BLUP}}$ requires the linear model to be valid.

Remark: In practice, auxiliary variables for which the population total is known are usually few in number and not sufficiently predictive of the variable y for eliminating the selection bias. These may be supplemented with other auxiliary variables for which the total can be estimated using an existing probability survey. Therefore, the vector of population totals may be a blend of known and estimated totals. If the probability survey itself is calibrated to known population totals, then only the estimated totals $\hat{\mathbf{T}}_{\mathbf{x}}$ from the probability survey can be used.

A linear model is not always appropriate. This is the case when the variable y is categorical. Another typical example occurs when it is desired to estimate the total of a quantitative variable in a domain of interest. The variable y is then defined as the product of that quantitative variable and a binary variable indicating domain membership. To model such a variable, it is natural to consider a mixture of a degenerate distribution at 0 and a continuous distribution. When the relationship between y_k and $\mathbf{x}_k$ is not linear, model-assisted calibration of Wu and Sitter (2001) can be used to preserve the weighted form of the predictor θ while taking into account the non-linearity of the relationship. Suppose that we replace the above linear model with a non-linear (or non-parametric) model such that $E(y_k \mid \mathbf{X}) = h(\mathbf{x}_k)$, where $h(\cdot)$ is some function. The Wu and Sitter (2001) calibration first involves predicting y_k by $\hat{y}_k = \hat{h}(\mathbf{x}_k),\, k \in U,$ where $\hat{h}(\mathbf{x}_k)$ is a model-based estimate of $h(\mathbf{x}_k)$. Then, the total $T_{\hat{y}} = \sum_{k \in U} \hat{y}_k$ is calculated, and weights, $w_k^{\text{MC}},\, k \in s_{\text{NP}},$ are found that satisfy the calibration equation:

$$\sum_{k \in s_{\text{NP}}} w_k^{\text{MC}} \begin{pmatrix} 1 \\ \hat{y}_k \end{pmatrix} = \begin{pmatrix} N \\ T_{\hat{y}} \end{pmatrix}.$$

In other words, the equation (4.2) can be used, where $\mathbf{x}_k'$ is replaced with $(1,\ \hat{y}_k)$. This method requires knowing the population size N as well as the vector $\mathbf{x}_k$ for all units in the population U. If N and $T_{\hat{y}}$ are unknown, they can be replaced with estimates from a probability survey. For example, we can replace N with $\hat{N} = \sum_{k \in s_P} w_k$ and $T_{\hat{y}}$ with $\hat{T}_{\hat{y}} = \sum_{k \in s_P} w_k \hat{y}_k$. The approach can also be extended to the case of multiple variables of interest.

We mentioned that the selection bias may be considerably reduced if $\mathbf{x}_k$ is rich and contains variables that are related to both δ_k and y_k, which makes assumption 3 more realistic. It can therefore be useful in practice to consider a large number of potential auxiliary variables and select the most relevant ones using a variable selection technique. Chen, Valliant and Elliott (2018) suggest the LASSO technique for selecting auxiliary variables and show its good properties.

It should be noted that the predictor $\hat{\theta}^{\text{BLUP}}$ reduces to the naive estimator, $\hat{\theta}^{\text{NP}}$, in the simplest case possible where only one constant auxiliary variable is used: $x_k = 1,\, k \in U$. The naive estimator is usually highly biased. Its bias can be significantly reduced if the population U can be subdivided into H disjoint and exhaustive post-strata, $U_h,\, h = 1, \ldots, H,$ of size N_h. The post-stratification model, $E(y_k \mid \mathbf{X}) = \beta_h,\, k \in U_h,$ is then postulated, which is an important special case of the above linear model. Assuming that the variance $\operatorname{var}(y_k \mid \mathbf{X})$ is constant for $k \in U_h$, the predictor $\hat{\theta}^{\text{BLUP}}$ is written: $\hat{\theta}^{\text{BLUP}} = \sum_{h=1}^{H} N_h \hat{\beta}_h,$ where $\hat{\beta}_h = \sum_{k \in s_{\text{NP},h}} y_k / n_h^{\text{NP}},\, s_{\text{NP},h}$ is the set of units in U_h that are part of the sample s_{NP} and n_h^{NP} is the size of $s_{\text{NP},h}$. If the population sizes N_h are unknown, they can be replaced with estimates, $\hat{N}_h = \sum_{k \in s_{P,h}} w_k$, from a probability survey, where $s_{P,h}$ is the set of units in U_h that are part of the sample s_P. Regression trees could prove to be an interesting approach for forming post-strata, especially when the auxiliary variables are categorical.

If multiple categorical auxiliary variables are available, it can be useful to form a large number of post-strata to reduce the selection bias. If many auxiliary variables are crossed, the sample sizes n_h^{NP} could become very small, thereby making the estimators $\hat{\beta}_h$ very unstable. Gelman and Little (1997) suggest using a multi-level regression model to obtain estimators $\tilde{\beta}_h$ more stable than $\hat{\beta}_h$. They then consider the post-stratified predictor: $\hat{\theta}^{\text{MRP}} = \sum_{h=1}^{H} N_h \tilde{\beta}_h$. Nowadays, this method is known as Mr.P or MRP (Multilevel Regression and Poststratification); see, for example, Mercer et al. (2017). A similar approach would use small area estimation methods (Rao and Molina, 2015) to stabilize the estimators $\hat{\beta}_h$. Although such methods are likely to produce much more precise estimates of the average of variable y over the population U_h, it remains to be determined whether such methods can produce significant efficiency gains for estimating the overall total θ compared to the simple post-stratified predictor $\hat{\theta}^{\text{BLUP}} = \sum_{h=1}^{H} N_h \hat{\beta}_h$. It seems that regression trees provide another way to control the instability of the estimators $\hat{\beta}_h$ since a criterion is generally used to prevent an overly narrow subdivision of the population. These various methods warrant further investigation in future research. Precise estimation of population sizes N_h, if not known, is also a problem not to be overlooked when the population is divided into a large number of post-strata.

4.2 Statistical matching

Statistical matching, or data fusion, is an approach developed for combining data from two different sources that contain both source-specific variables and common variables. Readers are referred to D'Orazio, Di Zio and Scanu (2006) or Rässler (2012) for a review of statistical matching methods. In the context of this article, statistical matching involves modelling the relationship between y_k and the auxiliary variables $\mathbf{x}_k$, which are common to both sources, using data from the non-probability sample. As with calibration, the non-probability sample selection mechanism is assumed to be non-informative, and the auxiliary variables must be chosen carefully in order to make assumption 3 as plausible as possible. Once a model has been determined, it is used to predict the y values in a probability sample. Statistical matching can be viewed as an imputation problem with an imputation rate of 100%. The predictor of θ, obtained from the probability sample, takes the form: $\hat{\theta}^{\text{SM}} = \sum_{k \in s_P} w_k y_k^{\text{imp}}$, where y_k^{imp} is the imputed value for the unit $k \in s_P$. As in calibration, inferences are conditional on $\boldsymbol{\delta}$ and $\mathbf{X}$. Assumption 3, in a statistical matching context, can be viewed as analogous to the Population Missing At Random (PMAR) assumption introduced by Berg, Kim and Skinner (2016) in a non-response context.

If the linear regression model $E(y_k \mid \mathbf{X}) = \mathbf{x}_k' \boldsymbol{\beta}$ is used, the imputed value for the unit $k \in s_P$ is $y_k^{\text{imp}} = \mathbf{x}_k' \hat{\boldsymbol{\beta}}$ and the resulting predictor is given by $\hat{\theta}^{\text{SM}} = \hat{\mathbf{T}}_{\mathbf{x}}' \hat{\boldsymbol{\beta}}$. If assumptions 1 to 3 are satisfied and $E(\hat{\mathbf{T}}_{\mathbf{x}} \mid \boldsymbol{\Omega}_P) = \mathbf{T}_{\mathbf{x}}$, statistical matching produces an unbiased predictor, $\hat{\theta}^{\text{SM}}$, i.e., $E(\hat{\theta}^{\text{SM}} - \theta \mid \boldsymbol{\delta}, \mathbf{X}) =$ 0. Also, if $v_k = \mathbf{x}_k' \boldsymbol{\lambda}$, for a certain known vector $\boldsymbol{\lambda}$, it can be shown that $\sum_{k \in s_{\text{NP}}} (y_k - \mathbf{x}_k' \hat{\boldsymbol{\beta}}) = 0$, and the predictor $\hat{\theta}^{\text{SM}}$ is equivalent to the predictor $\hat{\theta}^{\text{BLUP}}$ if we replace $\mathbf{T}_{\mathbf{x}}$ in (4.1) with $\hat{\mathbf{T}}_{\mathbf{x}}$. It can also be shown that, for a post-stratification model where we impute $y_k, k \in s_{P,h},$ with $y_k^{\text{imp}} = \hat{\beta}_h,$ the predictor $\hat{\theta}^{\text{SM}}$ reduces to $\hat{\theta}^{\text{SM}} = \sum_{h=1}^{H} \hat{N}_h \hat{\beta}_h.$ Therefore, statistical matching and calibration produce similar predictors, even identical in some cases, when a linear model is postulated and the totals $\mathbf{T}_{\mathbf{x}}$ are estimated.

Choosing between statistical matching or calibration can depend on the user's perspective. For example, if it is the content of the non-probability source, in terms of variables of interest, that is relevant to the user, then it seems natural to weight the non-probability sample in the hopes of reducing the selection bias for all variables of interest. The calibration technique or the methods in Section 4.3 are obvious choices for such weighting. Conversely, if instead it is the content of the probability survey that is relevant, then statistical matching is the appropriate choice. This method enriches the probability survey by imputing the missing variables of interest.

Statistical matching is easily generalized to non-linear or non-parametric models such that $E(y_k \mid \mathbf{X}) = h(\mathbf{x}_k)$. The imputed values y_k^{imp} are simply obtained by predicting the missing values $y_k,$ $k \in s_P,$ using the chosen model. The predictor $\hat{\theta}^{\text{SM}} = \sum_{k \in s_P} w_k y_k^{\text{imp}}$ remains unbiased if assumptions 1 to 3 are satisfied and if $E\left(y_k^{\text{imp}} - y_k \mid \boldsymbol{\delta}, \mathbf{X}\right) = 0.$ Donor or nearest neighbour imputation is a non-parametric imputation method commonly used for handling non-response (see, for example, Beaumont and Bocci, 2009) that does not require a linear relationship between y_k and $\mathbf{x}_k.$ In the context of matching non-probability and probability samples, donor imputation was popularized by Rivers (2007). For a given unit $k \in s_P,$ the method involves finding the nearest donor, with respect to the auxiliary variables $\mathbf{x}$, among the units of the non-probability sample and replacing the missing value y_k with the y value from this donor. For donor imputation, the condition $E\left(y_k^{\text{imp}} - y_k \mid \boldsymbol{\delta}, \mathbf{X}\right) = 0$ is satisfied if, for each recipient $k \in s_P,$ the donor has exactly the same values of $\mathbf{x}$ as the recipient. When one or more auxiliary variables are continuous, this condition is satisfied only asymptotically in general. A very large non-probability sample provides a large pool of donors, which should help to approximately satisfy this condition.

Remark: In some applications, a very large non-probability panel of volunteers, $s_{\text{NP}},$ is available, which contains a few auxiliary variables for matching, $\mathbf{x}$, but no variable of interest. Ideally, the variables of interest would be collected for all units of the panel $s_{\text{NP}},$ but that is impossible due to the cost and the burden on the panel members. Therefore, in practice, a sub-sample s_{NP}^* of s_{NP} is selected using random or non-random sampling methods. Quota sampling (e.g., Deville, 1991) is often considered in this context. In addition to collecting the variables of interest for all units of $s_{\text{NP}}^*,$ there may also be interest in collecting other auxiliary variables for matching in order to enhance the vector $\mathbf{x}$. The matching can then be done to the probability sample, often much smaller in size, as long as the latter contains the same auxiliary variables as those of the non-probability sub-sample $s_{\text{NP}}^*.$ By carefully choosing the auxiliary variables for the matching, the potential for bias reduction is increased (Schonlau and Cooper, 2017). The implementation proposed by Rivers (2007) is slightly different. Rivers (2007) suggests conducting the matching between the probability sample and the panel s_{NP} using the auxiliary variables available in both sources. The variables of interest are collected only for the set of donors in s_{NP} who have been matched to a unit in the probability sample, which allows for a significant reduction of data collection costs and burden. The implicit assumption is that the panel members, initially volunteers, are more likely to respond than individuals chosen at random in the population. Obviously, non-response is unavoidable, and this problem must be dealt with, potentially through imputation. The advantage of this method is that the matching is carried out using the panel s_{NP} rather than a sub-sample of this panel; the pool of donors is larger. However, the matching cannot be done using the enhanced vector of auxiliary variables because it is not available for the units in s_{NP}, which limits the potential for bias reduction.

Lavallée and Brisbane (2016) point out the connection between statistical matching and indirect sampling (Lavallée, 2007; Deville and Lavallée, 2006). They propose an estimator obtained by imputing each missing value $y_k, k \in s_P$, by a weighted average of the y values of nearest donors. In reality, their estimator can also be obtained equivalently by imputing the missing values using fractional donor imputation (for example, Kim and Fuller, 2004). The use of more than one donor to impute the missing values yields a typically modest variance reduction.

Several imputation methods used in practice can be considered linear (Beaumont and Bissonnette, 2011). This is the case for linear regression imputation, donor imputation and fractional donor imputation. An imputation method is said to be linear if the imputed value $y_k^{\text{imp}}, k \in s_P$, can be written as $y_k^{\text{imp}} = \sum_{l \in s_{\text{NP}}} \omega_{kl} y_l$, where ω_{kl} is a function of $\boldsymbol{\delta}$ or $\mathbf{X}$ but not of $\mathbf{Y}$. For example, for donor or nearest-neighbour imputation, $\omega_{kl} = 1$ if the unit $l \in s_{\text{NP}}$ is the donor for the recipient $k \in s_P$; otherwise $\omega_{kl} = 0$. For a linear imputation method, the estimator $\hat{\theta}^{\text{SM}} = \sum_{k \in s_P} w_k y_k^{\text{imp}}$ can be rewritten as a weighted sum over the non-probability sample: $\hat{\theta}^{\text{SM}} = \sum_{l \in s_{\text{NP}}} W_l y_l$, where $W_l = \sum_{k \in s_P} w_k \omega_{kl}$. Therefore, for linear imputation methods, statistical matching is an alternative to calibration and to the methods in Section 4.3 if the objective is to properly weight the non-probability sample.

So far, we have considered only the estimation of the total $\theta = \sum_{k \in U} y_k$. However, the probability sample contains other variables, and there may be interest in the relationship between two or more variables, some from the probability survey and others imputed from the non-probability sample. As an example, suppose that the estimation of the total $\theta = \sum_{k \in U} \tilde{y}_k y_k$ is of interest, where $\tilde{y}_k$ is a variable collected in the probability survey, but not available in the non-probability sample. It could, for example, define membership in a domain of interest. Statistical matching can be used to estimate this parameter by $\hat{\theta}^{\text{SM}} = \sum_{k \in s_P} w_k \tilde{y}_k y_k^{\text{imp}}$. We use $\tilde{\mathbf{Y}}$ to denote the vector that contains the values of the variable $\tilde{y}_k$, $k \in U$. It can be shown that $\hat{\theta}^{\text{SM}}$ is unbiased, $E\left(\hat{\theta}^{\text{SM}} - \theta \,\middle|\, \boldsymbol{\delta}, \mathbf{X}, \tilde{\mathbf{Y}}\right) = 0$, if assumptions 1 to 3 are satisfied in addition to the following assumption:

Assumption 4: $\mathbf{Y}$ and $\tilde{\mathbf{Y}}$ are independent after conditioning on $\boldsymbol{\delta}$ and $\mathbf{X}$.

Assumption 4 is known as the conditional independence assumption in the statistical matching literature.

4.3 Inverse propensity score weighting

Instead of modelling the relationship between y_k and $\mathbf{x}_k$, the relationship between δ_k and $\mathbf{x}_k$ could be modelled. The main advantage of this approach is to simplify the modelling effort when there are multiple variables of interest since there is always only one variable δ_k. With this approach, inferences are conditional on $\mathbf{Y}$ and $\mathbf{X}$. Also, it is usually assumed that assumption 3 is valid and thus $\Pr(\delta_k = 1 \mid \mathbf{Y}, \mathbf{X}) = \Pr(\delta_k = 1 \mid \mathbf{X})$. The probability of participation $p_k = \Pr(\delta_k = 1 \mid \mathbf{X})$ is then estimated by $\hat{p}_k$, and the estimate $\hat{\theta}^{\text{PS}} = \sum_{k \in s_{\text{NP}}} w_k^{\text{PS}} y_k$ is calculated, where $w_k^{\text{PS}} = 1 / \hat{p}_k$. The assumption that $p_k > 0, k \in U$, must be made. It is called the *positivity* assumption by Mercer et al. (2017). It may also be required in the calibration and statistical matching approaches. For example, empty post-strata $(n_h^{\text{NP}} = 0)$ may occur if it is not satisfied. To fix this issue, these empty post-strata are usually collapsed with other non-empty post-strata. This collapsing may jeopardize the validity of assumption 3 if the collapsed post-strata are different.

The estimation of p_k can be achieved by postulating a parametric model $p_k = g(\mathbf{x}_k; \boldsymbol{\alpha})$, where g is some function, normally bounded by 0 and 1, and $\boldsymbol{\alpha}$ is a vector of unknown model parameters. The logistic function $g(\mathbf{x}_k; \boldsymbol{\alpha}) = \exp(\mathbf{x}_k' \boldsymbol{\alpha}) \Big/ \left[1 + \exp(\mathbf{x}_k' \boldsymbol{\alpha})\right]$ predominates in the applications (see Kott, 2019, for a recent application). The estimator of $\boldsymbol{\alpha}$ is denoted by $\hat{\boldsymbol{\alpha}}$ and the estimated probability by $\hat{p}_k = g(\mathbf{x}_k; \hat{\boldsymbol{\alpha}})$. Ideally, $\boldsymbol{\alpha}$ would be estimated using $\mathbf{x}_k$ for all the units in the population U similar to what would be done in a non-response context. For example, assuming the logistic function is used, $\boldsymbol{\alpha}$ could be estimated by solving the maximum likelihood equation:

$$\sum_{k \in U} [\delta_k - p_k(\boldsymbol{\alpha})] \mathbf{x}_k = \sum_{k \in s_{\text{NP}}} \mathbf{x}_k - \sum_{k \in U} p_k(\boldsymbol{\alpha}) \mathbf{x}_k = \mathbf{0}. \tag{4.3}$$

This is impossible when $\mathbf{x}_k$ is not known for all units $k \in U - s_{\text{NP}}$, which is almost always the case in practice. Iannacchione, Milne and Folsom (1991) proposed another unbiased estimation equation for $\boldsymbol{\alpha}$ (see also Deville and Dupont, 1993):

$$\sum_{k \in s_{\text{NP}}} \frac{\mathbf{x}_k}{p_k(\boldsymbol{\alpha})} - \sum_{k \in U} \mathbf{x}_k = \mathbf{0}. \tag{4.4}$$

The main advantage of equation (4.4) is that it does not require knowing $\mathbf{x}_k$ for each unit $k \in U - s_{\text{NP}}$. However, it is necessary to have access to the vector of totals $\sum_{k \in U} \mathbf{x}_k$ from an external source. An interesting property of equation (4.4) is that the resulting weights $w_k^{\text{PS}} = 1 / p_k(\hat{\boldsymbol{\alpha}})$ satisfy the calibration equation $\sum_{k \in s_{\text{NP}}} w_k^{\text{PS}} \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k$, just like the weights w_k^C given in (4.2). Indeed, it can be shown that solving (4.4) yields $w_k^{\text{PS}} = w_k^C$ if the model $p_k(\boldsymbol{\alpha}) = \left(1 + v_k^{-1} \mathbf{x}_k' \boldsymbol{\alpha}\right)^{-1}$ is used. However, this is a less natural model than the above logistic model for modelling a probability.

To get around the problem of missing values $\mathbf{x}_k, k \in U - s_{\text{NP}}$, Chen et al. (2019) suggest estimating $\sum_{k \in U} p_k(\boldsymbol{\alpha}) \mathbf{x}_k$ in (4.3) using a probability survey. The equation to be solved becomes:

$$\sum_{k \in s_{\text{NP}}} \mathbf{x}_k - \sum_{k \in s_P} w_k p_k(\boldsymbol{\alpha}) \mathbf{x}_k = \mathbf{0}. \tag{4.5}$$

Equation (4.5) is unbiased conditionally on $\mathbf{Y}$ and $\mathbf{X}$ provided that the probability survey allows for unbiased estimation, conditionally on $\mathbf{Y}$ and $\mathbf{\Omega}$, of any population total that is not a function of $\boldsymbol{\delta}$ such as $\sum_{k \in U} p_k(\boldsymbol{\alpha})\,\mathbf{x}_k$. Assumptions 1 and 3 are required, but not assumption 2. Using the idea of Iannacchione et al. (1991), an alternative to (4.5) is obtained by solving:

$$\sum_{k \in s_{\text{NP}}} \frac{\mathbf{x}_k}{p_k(\boldsymbol{\alpha})} - \sum_{k \in s_P} w_k \mathbf{x}_k = \mathbf{0}. \tag{4.6}$$

Equation (4.6) produces weights $w_k^{\text{PS}} = 1 / p_k(\hat{\boldsymbol{\alpha}})$ that satisfy the calibration equation $\sum_{k \in s_{\text{NP}}} w_k^{\text{PS}} \mathbf{x}_k = \sum_{k \in s_P} w_k \mathbf{x}_k$ (see also Lesage, 2017; Rao, 2020). The estimators of $\boldsymbol{\alpha}$ obtained using (4.5) or (4.6) are likely less efficient than those obtained using (4.3) or (4.4). If $\mathbf{x}_k$, $k \in U - s_{\text{NP}}$, or the vector $\sum_{k \in U} \mathbf{x}_k$ is known, then using (4.3) or (4.4) is preferable. Otherwise, the estimating equations (4.5) or (4.6) can be used provided that $\mathbf{x}_k$ is collected in a probability survey. Note that the indicators δ_k do not need to be observed in the probability sample.

Equations (4.5) and (4.6) may be more difficult to solve than equations (4.3) and (4.4) and may not have a solution. Consider, for example, the case where there is only one auxiliary variable: $x_k = 1$. Using (4.5) or (4.6), it can be seen that the estimated probability reduces to: $\hat{p}_k = n^{\text{NP}} \big/ \sum_{k \in s_P} w_k$. If the size of the probability sample is sufficiently large, it is expected that $0 < \hat{p}_k < 1$. For small sample sizes, it may happen that $\hat{p}_k > 1$ due to the variability of $\sum_{k \in s_P} w_k$. In that case, equations (4.5) and (4.6) would not have a solution if the logistic function is used since it requires that $0 < \hat{p}_k < 1$. To avoid this issue, it may be helpful to consider other functions not bounded by 1, such as $g(\mathbf{x}_k; \boldsymbol{\alpha}) = \exp(\mathbf{x}_k' \boldsymbol{\alpha})$.

Kim and Wang (2019) suggest using the probability sample to estimate the participation probability. Assuming the logistic function is used, the equation to be solved is:

$$\sum_{k \in s_P} w_k [\delta_k - p_k(\boldsymbol{\alpha})]\,\mathbf{x}_k = \sum_{k \in s_P} w_k \delta_k \mathbf{x}_k - \sum_{k \in s_P} w_k p_k(\boldsymbol{\alpha})\,\mathbf{x}_k = \mathbf{0}.$$

The method requires knowing the indicators δ_k in the probability sample and the validity of assumptions 1, 2 and 3 to ensure the estimating equation is unbiased. Also, the probability sample size is usually small relative to the non-probability sample size, and it can be numerically difficult to estimate $\boldsymbol{\alpha}$, especially when $\mathbf{x}_k$ contains a large number of variables and the overlap between the two samples is small.

Lee (2006), see also Rivers (2007), Valliant and Dever (2011) and Elliott and Valliant (2017), proposes to combine the two samples and then estimate p_k using logistic regression. It seems that the author implicitly assumes that the two samples do not overlap, i.e., that $\delta_k = 0$ for all units in s_P. Using again the logistic function, the resulting estimating equation is:

$$\sum_{k \in s_{\text{NP}}} \eta_k^{\text{NP}} [1 - p_k(\boldsymbol{\alpha})]\,\mathbf{x}_k - \sum_{k \in s_P} w_k p_k(\boldsymbol{\alpha})\,\mathbf{x}_k = \mathbf{0}, \tag{4.7}$$

where η_k^{NP} is a certain weight for the units in the non-probability sample. The method is somewhat similar to the one proposed by Chen et al. (2019), but the estimating equation (4.7) is not unbiased, conditionally on $\mathbf{Y}$ and $\mathbf{X}$, unlike equations (4.5) and (4.6). However, if we assume $\eta_k^{\text{NP}} = 1$ and if $\max\{p_k;\ k \in U\}$ is small, equation (4.7) becomes approximately equivalent to equation (4.5). Yet Lee (2006) does not directly use the estimated probabilities resulting from (4.7). The author uses them only to order the union of the two samples and then create homogeneous classes. Using homogeneous classes brings some robustness to model misspecification and can help prevent very small estimated probabilities and thus very large weights. In the context of non-response, forming homogeneous imputation or reweighting classes was studied by Little (1986), Eltinge and Yansaneh (1997), and Haziza and Beaumont (2007), among others. Haziza and Lesage (2016) illustrate the robustness of the method when the function $g(\mathbf{x}_k;\ \boldsymbol{\alpha})$ is misspecified. The method is used regularly in Statistics Canada surveys for dealing with non-response.

Rather than using (4.7), homogeneous classes could be formed by starting with the unbiased equations (4.5) or (4.6). These initial estimated probabilities are denoted by $\hat{p}_k^0 = g(\mathbf{x}_k;\ \hat{\boldsymbol{\alpha}})$. The sample $s = s_P \cup s_{\text{NP}}$ can then be sorted by $\hat{p}_k^0$ and divided into C homogeneous classes of equal or unequal sizes. The set of units in s_P that are part of class c is denoted by $s_{P,\,c}$ whereas the set of units in s_{NP} that are part of class c is denoted by $s_{\text{NP},c}$. The weight w_k^{PS} for a unit $k \in s_{\text{NP},c}$ is equal to the inverse of the estimated participation rate in class c and is given by $w_k^{\text{PS}} = \hat{N}_c \big/ n_c^{\text{NP}}$, where $\hat{N}_c = \sum_{k \in s_{P,c}} w_k$ and n_c^{NP} is the number of units in $s_{\text{NP},c}$ This weight ensures the calibration property: $\sum_{k \in s_{\text{NP},c}} w_k^{\text{PS}} = \hat{N}_c$. The number of classes must be large enough to capture a high percentage of the variability of the initial probabilities $\hat{p}_k^0$, thereby reducing the bias. On the other hand, it must not be too large to prevent the occurrence of empty classes since the weights $w_k^{\text{PS}} = \hat{N}_c \big/ n_c^{\text{NP}}$ cannot be calculated if $n_c^{\text{NP}} = 0$. Regression trees can prove to be an effective alternative for forming classes. In a non-response context, they have been studied by Phipps and Toth (2012). The estimator $\hat{\theta}^{\text{PS}} = \sum_{k \in s_{\text{NP}}} w_k^{\text{PS}} y_k$ obtained after forming homogeneous classes has exactly the same form as the post-stratified estimator described in the calibration approach in Section 4.1; the only difference is that the classes are built by modelling δ_k rather than y_k.

Assumption 3 may not be realistic in some contexts so that $\Pr(\delta_k = 1 \mid \mathbf{Y}, \mathbf{X}) \neq \Pr(\delta_k = 1 \mid \mathbf{X})$. In this case, the participation probability $p_k = \Pr(\delta_k = 1 \mid \mathbf{Y}, \mathbf{X})$ might be modelled using a vector of explanatory variables $\mathbf{x}_k^*$, defined using the variable of interest y_k (or variables of interest if there are several) and potentially other auxiliary variables $\mathbf{x}_k$. A parametric model, $p_k = g(\mathbf{x}_k^*;\ \boldsymbol{\alpha})$, can be considered for modelling the participation probability. Equations (4.5) and (4.6) cannot be used to estimate $\boldsymbol{\alpha}$ because y_k (and therefore $\mathbf{x}_k^*$) is not available in the probability sample. However, an equation similar to (4.6) can be used:

$$\sum_{k \in s_{\text{NP}}} \frac{\mathbf{x}_k^I}{g(\mathbf{x}_k^*;\ \boldsymbol{\alpha})} - \sum_{k \in s_P} w_k \mathbf{x}_k^I = \mathbf{0}. \tag{4.8}$$

The vector $\mathbf{x}_k^I$, of the same size as $\boldsymbol{\alpha}$, contains calibration variables, also called instrumental variables in the econometric literature. We use $\mathbf{X}^I$ to denote the matrix that contains the values of vector $\mathbf{x}_k^I$, $k \in U$. Equation (4.8) requires knowing the calibration variables $\mathbf{x}_k^I$ for both samples. However, the explanatory variables $\mathbf{x}_k^*$ can be observed only for the units in the non-probability sample. Equation (4.8) produces weights $w_k^{\text{PS}} = 1 / g(\mathbf{x}_k^*;\ \hat{\boldsymbol{\alpha}})$ that satisfy the calibration equation $\sum_{k \in s_{\text{NP}}} w_k^{\text{PS}} \mathbf{x}_k^I = \sum_{k \in s_P} w_k \mathbf{x}_k^I$. An equation similar to (4.8) was originally proposed by Deville (1998) to deal with non-response (see also Kott, 2006; Haziza and Beaumont, 2017). Equation (4.8) is unbiased, conditionally on $\mathbf{Y}$, $\mathbf{X}$ and $\mathbf{X}^I$, if the instrumental variables $\mathbf{x}_k^I$ can be selected such that the following assumption is satisfied:

Assumption 5: $\boldsymbol{\delta}$ and $\mathbf{X}^I$ are independent after conditioning on $\mathbf{Y}$ and $\mathbf{X}$.

Assumption 3 is no longer required, but is replaced with another assumption. The choice of instrumental variables $\mathbf{x}_k^I$ that satisfy assumption 5 is not always obvious in practice. They must not be predictive of δ_k after conditioning on $\mathbf{x}_k^*$. Ideally, for efficiency reasons, the instrumental variables are selected so as to be predictive of $\mathbf{x}_k^*$ without compromising assumption 5. Unlike equations (4.5) and (4.6), equation (4.8) cannot be used to form homogeneous classes because the participation probabilities $\hat{p}_k = g(\mathbf{x}_k^*;\ \hat{\boldsymbol{\alpha}})$ cannot be calculated for the units in the probability sample. As such, the property of robustness that comes with homogeneous classes is lost. Because of these drawbacks, equation (4.8) should be considered only when there are strong reasons to believe that assumption 3 is not appropriate.

Once weights w_k^{PS} have been calculated using one of the methods in this section, they can still be adjusted through calibration. The objective of this calibration is to improve the precision of the estimator $\hat{\theta}^{\text{PS}}$ and also obtain a double robustness property (see Chen et al., 2019).

In general, the variable y is observed for the entire non-probability sample, and the inverse propensity-score weighted estimator, $\hat{\theta}^{\text{PS}} = \sum_{k \in s_{\text{NP}}} w_k^{\text{PS}} y_k$, or a weighted estimator obtained by calibration or statistical matching can be used. Sometimes, the non-probability sample is too large and the variable y can only be collected for a sub-sample of s_{NP}. Quota sampling (e.g., Deville, 1991) is a commonly used method for drawing the sub-sample if auxiliary variables are available for $k \in s_{\text{NP}}$. An alternative to quota sampling is to calculate the weights w_k^{PS} for the entire non-probability sample and use them to select a random sub-sample with probabilities proportional to the weights. The variable y is then collected only for the sub-sample, and the estimates are obtained as if the sub-sample was drawn from the population using an equal probability design. This approach is called inverse sampling in the literature on probability surveys (see, for example, Hinkins, Oh and Scheuren, 1997; or Rao, Scott and Benhin, 2003) and was proposed by Kim and Wang (2019) for non-probability samples.

4.4 Small area estimation

In most surveys, it is desired to estimate the total of the variable y not just for the entire population U, but also for different subgroups of the population, called domains. Probability surveys conducted by national statistical agencies generally produce reliable estimates for domains with a sufficient number of sample units. Their bias is controlled through the various sampling and data collection procedures, and their variance is typically small enough to draw accurate conclusions. When the domain of interest contains few sample units, the survey estimates may become unstable to the point of being unusable even when their bias stays under control. To remedy a lack of data in a domain of interest, small area estimation methods may be considered. These methods offset the lack of observed data in a domain through model assumptions that link auxiliary data to survey data. Two types of models are commonly used: unit-level models and area-level models. The area-level model of Fay and Herriot (1979) is undoubtedly the most popular. It requires auxiliary data to be available at the domain level only, unlike unit-level models, which require auxiliary variables for each unit of the population U. Readers are referred to Rao and Molina (2015) for an excellent coverage of the various approaches. Below, we focus on the Fay-Herriot model.

Suppose it is desired to estimate D totals, $\theta_d = \sum_{k \in U_d} y_k$, $d = 1, \ldots, D$, where U_d are D disjoint subsets of the population. Using a probability survey, θ_d can be estimated by $\hat{\theta}_d = \sum_{k \in s_{P,d}} w_k y_k$, where $s_{P,d}$ is the set of sample units that fall within domain d. The estimator $\hat{\theta}_d$ is called the direct estimator of θ_d because it only uses y values of units belonging to domain d. Small area estimation techniques generally lead to indirect estimators that combine the sample y values of domain d with y values of units outside domain d. We assume that a vector of auxiliary variables is available at the area level, and these variables come from sources independent of the probability sample. This vector for domain d is denoted by $\mathbf{x}_d$. For example, the vector $\mathbf{x}'_d = \left(N_d,\ N_d \hat{\mu}_d^{\mathrm{NP}}\right)$ could be considered, where N_d is the population size in domain d, $\hat{\mu}_d^{\mathrm{NP}} = \sum_{k \in s_{\mathrm{NP},d}} y_k^* \,/\, n_d^{\mathrm{NP}}$ is the average of variable y^* in a non-probability sample, $s_{\mathrm{NP},d}$ is the set of units in the non-probability sample that are in domain d and n_d^{NP} is the size of the non-probability sample in domain d. If the population size N_d is unknown, it can be replaced with an estimate independent of the probability survey. We use $\mathbf{X}$ to denote the matrix that contains the values of vector $\mathbf{x}_d$, $d = 1, \ldots, D$. Note that the vector $\boldsymbol{\delta}$ is hidden in the matrix $\mathbf{X}$ in this section.

The Fay-Herriot model has two components: the sampling model and the linking model. The sampling model is based on the assumption that, conditionally on $\boldsymbol{\Omega}_P$, the direct estimators $\hat{\theta}_d$ are independent and unbiased, i.e., $E\left(\hat{\theta}_d \mid \boldsymbol{\Omega}_P\right) = \theta_d$. Their design variance is denoted by $\psi_d = \operatorname{var}\left(\hat{\theta}_d \mid \boldsymbol{\Omega}_P\right)$. The sampling model is usually written in the form:

$$\hat{\theta}_d = \theta_d + e_d, \tag{4.9}$$

where e_d is the sampling error such that $E(e_d \mid \boldsymbol{\Omega}_P) = 0$ and $\operatorname{var}(e_d \mid \boldsymbol{\Omega}_P) = \psi_d$. The independence assumption of the estimators $\hat{\theta}_d$ (and therefore of the sampling errors e_d) can be questioned when the strata do not coincide with the domains of interest. Section 8.2 of Rao and Molina (2015) discusses methods that take into account correlated sampling errors. In practice, it is often assumed that these correlations are weak, and they are ignored.

The linking model assumes that, conditionally on $\mathbf{X}$, the totals θ_d are independent, $E(\theta_d \mid \mathbf{X}) = \mathbf{x}'_d\boldsymbol{\beta}$ and $\mathrm{var}(\theta_d \mid \mathbf{X}) = b_d^2\sigma_v^2$, where b_d are known constants used for controlling heteroscedasticity and $\boldsymbol{\beta}$ and σ_v^2 are unknown model parameters. The linking model is usually written in the form:

$$\theta_d = \mathbf{x}'_d\boldsymbol{\beta} + b_d v_d, \tag{4.10}$$

where v_d is the model error such that $E(v_d \mid \mathbf{X}) = 0$ and $\mathrm{var}(v_d \mid \mathbf{X}) = \sigma_v^2$. When the parameters of interest, θ_d, are totals, it is often appropriate to let $b_d = N_d$. From (4.9) and (4.10), we obtain the combined model:

$$\hat{\theta}_d = \mathbf{x}'_d\boldsymbol{\beta} + a_d, \tag{4.11}$$

where $a_d = b_d v_d + e_d$ is the combined error. When using the Fay-Herriot model (4.11), inferences are usually made conditionally on $\mathbf{X}$. It can easily be shown that $E(a_d \mid \mathbf{X}) = 0$ and $\mathrm{var}(a_d \mid \mathbf{X}) = b_d^2\sigma_v^2 + \tilde{\psi}_d$, where $\tilde{\psi}_d = E(\psi_d \mid \mathbf{X})$ is called the smooth design variance (Beaumont and Bocci, 2016; and Hidiroglou, Beaumont and Yung, 2019).

Now suppose that it is desired to predict the total θ_d using a linear predictor $\hat{\theta}_d^{\mathrm{LIN}} = \sum_{i=1}^{D} \lambda_{di}\hat{\theta}_i$, where λ_{di} are constants to be determined. A linear predictor uses all the data from the probability sample for predicting θ_d, not just the data from domain d. This explains how it derives its efficiency. However, not all linear predictors are appropriate for predicting θ_d. A strategy often used for determining the constants λ_{di} is to minimize the variance of the prediction error, $\mathrm{var}\left(\hat{\theta}_d^{\mathrm{LIN}} - \theta_d \,\middle|\, \mathbf{X}\right)$, subject to the constraint that the predictor must be unbiased, $E\left(\hat{\theta}_d^{\mathrm{LIN}} - \theta_d \,\middle|\, \mathbf{X}\right) = 0$. The resulting predictor, called the Best Linear Unbiased Predictor (BLUP), is denoted by $\hat{\theta}_d^{\mathrm{BLUP}}$, and can be written in the form (see, for example, Rao and Molina, 2015):

$$\hat{\theta}_d^{\mathrm{BLUP}} = \gamma_d\hat{\theta}_d + (1 - \gamma_d)\,\mathbf{x}'_d\hat{\boldsymbol{\beta}}, \tag{4.12}$$

where $\gamma_d = b_d^2\sigma_v^2 \big/ \left(b_d^2\sigma_v^2 + \tilde{\psi}_d\right)$ is bounded by 0 and 1, and

$$\hat{\boldsymbol{\beta}} = \left(\sum_{d=1}^{D} \frac{\mathbf{x}_d\mathbf{x}'_d}{b_d^2\sigma_v^2 + \tilde{\psi}_d}\right)^{-1} \sum_{d=1}^{D} \frac{\mathbf{x}_d}{b_d^2\sigma_v^2 + \tilde{\psi}_d}\,\hat{\theta}_d.$$

The predictor (4.12) is a weighted average of the direct estimator $\hat{\theta}_d$ and a prediction, $\mathbf{x}'_d\hat{\boldsymbol{\beta}}$, often called the synthetic estimator. More weight is given to the direct estimator when the smooth design variance, $\tilde{\psi}_d$, is small relative to the variance of the linking model, $b_d^2\sigma_v^2$. The predictor $\hat{\theta}_d^{\mathrm{BLUP}}$ is then similar to the direct estimator. This situation normally occurs when the sample size in the domain is large. Conversely, if the direct estimator is unstable and has a large smooth design variance, more weight is given to the synthetic estimator. If the number of domains is large, the prediction variance of $\hat{\theta}_d^{\mathrm{BLUP}}$, $\mathrm{var}\left(\hat{\theta}_d^{\mathrm{BLUP}} - \theta_d \,\middle|\, \mathbf{X}\right)$, is approximately equal to $\gamma_d\tilde{\psi}_d$. Since $\mathrm{var}\left(\hat{\theta}_d - \theta_d \,\middle|\, \mathbf{X}\right) = \tilde{\psi}_d$, the constant γ_d can be interpreted as being a variance reduction factor resulting from using $\hat{\theta}_d^{\mathrm{BLUP}}$ instead of $\hat{\theta}_d$. Therefore, the variance reduction is greater when γ_d is small, i.e., when the direct estimator is not precise. On the other hand, if the linking model is not properly specified, there is greater risk of significant bias when γ_d is small. To better understand this point, suppose that the real linking model is such that $E(\theta_d \mid \mathbf{X}) = \mu(\mathbf{x}_d)$ for some function $\mu(\cdot)$. Under this model, it can be shown that the bias of the predictor $\hat{\theta}_d^{\text{BLUP}}$ is given by

$$E\left(\hat{\theta}_d^{\text{BLUP}} - \theta_d \,\middle|\, \mathbf{X}\right) = -(1 - \gamma_d)\left(\mu(\mathbf{x}_d) - \mathbf{x}_d'\boldsymbol{\beta}_0\right), \tag{4.13}$$

where

$$\boldsymbol{\beta}_0 = \left(\sum_{d=1}^{D} \frac{\mathbf{x}_d \mathbf{x}_d'}{b_d^2 \sigma_v^2 + \tilde{\psi}_d}\right)^{-1} \sum_{d=1}^{D} \frac{\mathbf{x}_d}{b_d^2 \sigma_v^2 + \tilde{\psi}_d} \mu(\mathbf{x}_d).$$

If the linear model $\mu(\mathbf{x}_d) = \mathbf{x}_d'\boldsymbol{\beta}$ is valid, the bias disappears. Otherwise, the bias is not zero and increases as γ_d decreases or as the specification error of the linking model, $\mu(\mathbf{x}_d) - \mathbf{x}_d'\boldsymbol{\beta}_0$, increases. When γ_d is close to 1, the bias is usually negligible, but so is the variance reduction.

Remark: Note that the predictor $\hat{\theta}_d^{\text{BLUP}}$ and the bias (4.13) depend on the variance σ_v^2. If the linear model (4.10) is not valid, the parameters $\boldsymbol{\beta}$ and σ_v^2 no longer exist. Yet, the linking model (4.10) can still be postulated and its parameters can be estimated from the observed data as if the model were valid. The model variance σ_v^2, which enters in the calculation of the predictor $\hat{\theta}_d^{\text{BLUP}}$ and the bias (4.13), can be viewed as being the value towards which an estimator of σ_v^2 converges.

The predictor (4.12) cannot be calculated because it depends on the unknown variances σ_v^2 and $\tilde{\psi}_d$. When σ_v^2 and $\tilde{\psi}_d$ in (4.12) are replaced with estimators $\hat{\sigma}_v^2$ and $\hat{\tilde{\psi}}_d$, the BLUP (4.12) becomes the empirical best linear unbiased predictor, denoted as $\hat{\theta}_d^{\text{EBLUP}}$. There are a number of methods for estimating σ_v^2 (see Rao and Molina, 2015). One of the most commonly used methods is restricted maximum likelihood. To estimate $\tilde{\psi}_d$, we assume that a design-unbiased estimator of ψ_d is available, denoted by $\hat{\psi}_d$. This assumption is formally written: $E(\hat{\psi}_d \mid \boldsymbol{\Omega}_P) = \psi_d$. It follows that $E(\hat{\psi}_d \mid \mathbf{X}) = \tilde{\psi}_d$. Therefore, the estimator $\hat{\psi}_d$ is unbiased for $\tilde{\psi}_d$, but can be very unstable when the domain sample size is small. A more efficient approach for estimating $\tilde{\psi}_d$ involves modelling $\hat{\psi}_d$ given the auxiliary variables $\mathbf{x}_d$. In practice, a linear model is often used for $\log(\hat{\psi}_d)$, and it is assumed that the model errors follow a normal distribution (for example, Rivest and Belmonte, 2000). Beaumont and Bocci (2016), see also Hidiroglou et al. (2019), provide a method of moments for estimating $\tilde{\psi}_d$ that does not require the normality assumption.

The Fay-Herriot model requires the availability of auxiliary data only at the domain level. The variable y must be measured without error in the probability survey, but it is not essential for the auxiliary source to be perfect. This leaves the door open to all kinds of files external to the probability survey such as big data files. Kim, Wang, Zhu and Cruze (2018) is a recent example where an extension of the Fay-Herriot model was used with auxiliary data from satellite images. Small area estimation methods often achieve significant and sometimes impressive variance reductions (see, for example, Hidiroglou et al., 2019). The trade-off for obtaining these gains is the introduction of model assumptions and the risk that these assumptions do not hold. Therefore, model validation is a critical step in producing small area estimates, as in any model-based approach.

Small area estimation methods are generally used to improve the efficiency of estimators for domains with a small sample size. They could also be used to reduce the data collection costs and respondent burden by reducing the overall sample size of a probability survey for a few, if not all, survey variables. The estimates obtained from the reduced sample and the Fay-Herriot model, for example, could thus have a precision similar to the direct estimates from the probability survey obtained from the full sample. In this context, small area estimation methods would not be used to improve the precision for domains containing few units, but instead to reduce the overall data collection effort while preserving the quality of the estimates.

5 Conclusion

In this paper, we presented several methods that use data from a non-probability source while preserving a statistical framework that allows for valid inferences. This, in our view, is essential for national statistical agencies because, without this framework, the usual measures of the quality of the estimates, such as variance or mean square error estimates, disappear and it becomes difficult to draw accurate conclusions. Using data from a non-probability source is not without risk. For model-based approaches, it seems unavoidable to plan enough time and resources for modelling. The literature on classical statistics is replete with tools for validating model assumptions. Although this topic was not adequately covered in the previous sections, careful validation of the assumptions is still a critical step in the success of these approaches (Chambers, 2014) and is one of the recommendations made by Baker et al. (2013).

Estimating the variance or mean square error of the estimators described in the previous sections is also an important topic that we omitted. Yet, this problem does not pose any particular difficulties, in general, and a number of methods exist for variance or mean square error estimation. For design-based approaches, the topic has been extensively covered in the literature (see, for example, Wolter, 2007). This is also true for small area estimation methods (see Rao and Molina, 2015) and for the calibration approach (see Valliant, Dorfman and Royall, 2000). Nevertheless, it might be useful that research be undertaken to adequately address this issue in some specific cases, such as weighting by inverse propensity score or statistical matching by nearest donor.

We assumed that the non-probability source was a subset of the population of interest and that it may be subject to measurement errors. However, there are other potential flaws with non-probability sources. For example, they may contain duplicates or units outside the population. This could make some of the methods discussed in this article unusable, especially the design-based methods. Therefore, it might be useful to tackle these problems in the future.

We mainly limited ourselves to describing several methods that use data from a non-probability sample, whether or not combined with data from a probability survey, once all the data have been collected and processed. There are a number of other methods that use data from non-probability sources during the various stages of a probability survey. For example, one or more non-probability sources can be used to create a sampling frame or improve its coverage. These sources can also be used in a multi-frame sampling context, to replace data collection for certain variables, or to impute the missing values in a probability survey. These topics were not covered in this article, but are reviewed in Lohr and Raghunathan (2017).

The literature on integrating data of a probability and non-probability sample is quite recent. However, there are a number of methods that combine data from two probability surveys (e.g., Hidiroglou, 2001; Merkouris, 2004; Ybarra and Lohr, 2008; Merkouris, 2010; and Kim and Rao, 2012). Such methods may be used to first combine two probability surveys before integrating them with a non-probability source using one of the methods in Section 4. For example, if the total $\mathbf{T}_{\mathbf{x}}$ is unknown, it may be possible to estimate it using more than one probability survey and then use this estimated total in the calibration approach. It still needs to be assessed whether such a strategy would yield significant efficiency gains.

Are probability surveys bound to disappear for the production of official statistics? The question is relevant in the current context of surveys conducted by national statistical agencies where high data collection costs and increasingly lower response rates are observed. In our opinion, the time has not yet come because the alternatives are not reliable and general enough to eliminate the use of probability surveys without severely sacrificing the quality of the estimates. In Section 4, we mentioned that calibration and weighting by inverse propensity score could eliminate the use of a probability survey, provided that a vector of population totals $\mathbf{T}_{\mathbf{x}}$ is available from a census or a comprehensive administrative source. In general, these known totals will not be numerous and effective enough to sufficiently reduce the selection bias of a non-probability sample. To get around this problem, the suggestion has been made in the literature to complement $\mathbf{T}_{\mathbf{x}}$ with other totals estimated using a good-quality probability survey. It seems to us that this is the way to significantly reduce bias and to really take advantage of calibration and weighting by inverse propensity score methods presented in Section 4. Of course, some probability surveys with very low response rates and/or data of questionable quality could occasionally be eliminated in favour of data from non-probability sources. In our view, most surveys conducted by Statistics Canada do not fall into this category. Although they are not perfect, they continue to provide reliable information to meet users' needs and to make informed decisions. The complete elimination of probability surveys seems highly unlikely in the short or medium term. However, it can be expected that their use will be reduced in the future in order to control costs and respondent burden.

References

Baker, R., Brick, J.M., Bates, N.A., Battaglia, M., Couper, M.P., Dever, J.A., Gile, K. and Tourangeau, R. (2013). Summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology*, 1, 90-143.

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

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/en/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., and Bocci, C. (2016). *Small Area Estimation in the Labour Force Survey*. Paper presented at the Advisory Committee on Statistical Methods, May 2016, Statistics Canada.

Beaumont, J.-F., Bocci, C. and Hidiroglou, M. (2014). On weighting late respondents when a follow-up subsample of nonrespondents is taken. Paper presented at the Advisory Committee on Statistical Methods, May 2014, Statistics Canada.

Beaumont, J.-F., Haziza, D. and Bocci, C. (2014). An adaptive data collection procedure for call prioritization. *Journal of Official Statistics*, 30, 607-621.

Berg, E., Kim, J.-K. and Skinner, C. (2016). Imputation under informative sampling. *Journal of Survey Statistics and Methodology*, 4, 436-462.

Bethlehem, J. (2009). The rise of survey sampling. Discussion paper (09015), Statistics Netherlands, The Hague.

Bethlehem, J. (2016). Solving the nonresponse problem with sample matching. *Social Science Computer Review*, 34, 59-77.

Brick, J.M. (2011). The future of survey sampling. *Public Opinion Quarterly*, 75, 872-888.

Chambers, R. (2014). Survey sampling in official statistics – Some thoughts on directions. *Proceedings of the 2014 International Methodology Symposium*, Statistics Canada, Ottawa, Canada.

Chen, Y., Li, P. and Wu, C. (2019). Doubly robust inference with non-probability survey samples. *Journal of the American Statistical Association* (published online).

Chen, J.K.T., Valliant, R.L. and Elliott, M.R. (2018). Model-assisted calibration of non-probability sample survey data using adaptive LASSO. *Survey Methodology*, 44, 1, 117-144. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2018001/article/54963-eng.pdf.

Citro, C.F. (2014). From multiple modes for surveys to multiple data sources for estimates. *Survey Methodology*, 40, 2, 137-161. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2014002/article/14128-eng.pdf.

Couper, M.P. (2000). Web surveys: A review of issues and approaches. *Public Opinion Quarterly*, 64, 464-494.

Couper, M.P. (2013). Is the sky falling? New technology, changing media, and the future of surveys. *Survey Research Methods*, 7, 145-156.

Deville, J.-C. (1991). A theory of quota surveys. *Survey Methodology*, 17, 2, 163-181. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1991002/article/14504-eng.pdf.

Deville, J.-C. (1998). La correction de la non-réponse par calage ou par échantillonnage équilibré. *Proceedings of the Survey Methods Section*, Statistical Society of Canada, Sherbrooke, Canada.

Deville, J.-C., and Dupont, F. (1993). Non-réponse : principes et méthodes. *Actes des Journées de Méthodologie Statistique*, 53-69, December 15 and 16, 1993, INSEE, Paris.

Deville, J.-C., and Lavallée, P. (2006). Indirect sampling: The foundations of the generalized weight share method. *Survey Methodology*, 32, 2, 165-176. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2006002/article/9551-eng.pdf.

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

D'Orazio, M., Di Zio, M. and Scanu, M. (2006). *Statistical Matching: Theory and Practice*. New York: John Wiley & Sons, Inc.

Dutwin, D., and Buskirk, T.D. (2017). Apples to oranges or gala versus golden delicious? Comparing data quality of nonprobability internet samples to low response rate probability samples. *Public Opinion Quarterly*, 81, 213-249.

Elliott, M., and Valliant, R. (2017). Inference for non-probability samples. *Statistical Science*, 32, 249-264.

Eltinge, J.L., and Yansaneh, I.S. (1997). Diagnostics for formation of nonresponse adjustment cells, with an application to income nonresponse in the U.S. Consumer Expenditure Survey. *Survey Methodology*, 23, 1, 33-40. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1997001/article/3103-eng.pdf.

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., and Little, T.C. (1997). Poststratification into many categories using hierarchical logistic regression. *Survey Methodology*, 23, 2, 127-135. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1997002/article/3616-eng.pdf.

Haziza, D., and Beaumont, J.-F. (2007). On the construction of imputation classes in surveys. *International Statistical Review*, 75, 25-43.

Haziza, D., and Beaumont, J.-F. (2017). Construction of weights in surveys: A review. *Statistical Science*, 32, 206-226.

Haziza, D., and Lesage, É. (2016). A discussion of weighting procedures for unit nonresponse. *Journal of Official Statistics*, 32, 129-145.

Hidiroglou, M.A. (2001). Double sampling. *Survey Methodology*, 27, 2, 143-154. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2001002/article/6091-eng.pdf.

Hidiroglou, M.A., Beaumont, J.-F. and Yung, W. (2019). Development of a small area estimation system at Statistics Canada. *Survey Methodology*, 45, 1, 101-126. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2019001/article/00009-eng.pdf.

Hinkins, S., Oh, H.L. and Scheuren, F. (1997). Inverse sampling design algorithms. *Survey Methodology*, 23, 1, 11-22. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1997001/article/3101-eng.pdf.

Iannacchione, V.G., Milne, J.G. and Folsom, R.E. (1991). Response probability weight adjustments using logistic regression. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 637-642, Alexandria, VA.

Kalton, G. (2019). Developments in survey research over the past 60 years: A personal perspective. *International Statistical Review*, 87, S10-S30.

Kim, J.K., and Fuller, W. (2004). Fractional hot deck imputation. *Biometrika*, 91, 559-578.

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

Kim, J.K., and Tam, S.M. (2020). Data integration by combining big data and survey data for finite population inference. Unpublished manuscript.

Kim, J.K., and Wang, Z. (2019). Sampling techniques for big data analysis. *International Statistical Review*, 87, S177-S191.

Kim, J.K., Wang, Z., Zhu, Z. and Cruze, N.B. (2018). Combining survey and non-survey data for improved sub-area prediction using a multi-level model. *Journal of Agricultural, Biological and Environmental Statistics*, 23, 175-189.

Kott, P.S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, 32, 2, 133-142. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2006002/article/9547-eng.pdf.

Kott, P.S. (2019). A partially successful attempt to integrate a Web-recruited cohort into an address-based sample. *Survey Research Methods*, 13, 95-101.

Laflamme, F., and Karaganis, M. (2010). Development and implementation of responsive design for CATI surveys at Statistics Canada. *Proceedings of the European Conference on Quality in Official Statistics*, Helsinki, Finland, May 2010.

Lavallée, P. (2007). *Indirect Sampling*. New York: Springer.

Lavallée, P., and Brisbane, J. (2016). Sample matching: Towards a probabilistic approach for web surveys and big data? Paper presented at the Advisory Committee on Statistical Methods, May 2016, Statistics Canada.

Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel Web survey. *Journal of Official Statistics*, 22, 329-349.

Lesage, É. (2017). Combiner des données d'enquêtes probabilistes et des données massives non probabilistes pour estimer des paramètres de population finie. Unpublished manuscript.

Little, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 54, 139-157.

Lohr, S., and Raghunathan, T.E. (2017). Combining survey data with other data sources. *Statistical Science*, 32, 293-312.

Lundquist, P., and Särndal, C.-E. (2013). Aspects of responsive design with applications to the swedish living conditions survey. *Journal of Official Statistics*, 29, 557-582.

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *Annals of Applied Statistics*, 12, 685-726.

Mercer, A.W., Kreuter, F., Keeter, S. and Stuart, E.A. (2017). Theory and practice in nonprobability surveys: Parallels between causal inference and survey inference. *Public Opinion Quarterly*, 81, 250-271.

Merkouris, T. (2004). Combining independent regression estimators from multiple surveys. *Journal of the American Statistical Association*, 99, 1131-1139.

Merkouris, T. (2010). Combining information from multiple surveys by using regression for efficient small domain estimation. *Journal of the Royal Statistical Society: Series B*, 72, 27-48.

Miller, P.V. (2017). Is there a future for surveys? *Public Opinion Quarterly*, 81, 205-212.

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*, 97, 558-625.

Phipps, P., and Toth, D. (2012). Analyzing establishment nonresponse using an interpretable regression tree model with linked administrative data. *Annals of Applied Statistics*, 6, 772-794.

Rancourt, E. (2019). Admin-first as a statistical paradigm for Canadian statistics: Meaning, challenges and opportunities. *Proceedings of Statistics Canada's 2018 International Methodology Symposium* (to appear).

Rao, J.N.K. (2005). Interplay between sample survey theory and practice: An appraisal. *Survey Methodology*, 31, 2, 117-138. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2005002/article/9040-eng.pdf.

Rao, J.N.K. (2020). Making inference by combining data from multiple sources: An appraisal. *Sankhyā* (under review).

Rao, J.N.K., and Fuller, W. (2017). Sample survey theory and methods: Past, present and future directions. *Survey Methodology*, 43, 2, 145-160. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2017002/article/54888-eng.pdf.

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

Rao, J.N.K., Scott, A.J. and Benhin, E. (2003). Undoing complex survey data structures: Some theory and applications of inverse sampling (with discussion). *Survey Methodology*, 29, 2, 107-128. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2003002/article/6787-eng.pdf.

Rässler, S. (2012). *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*. Lecture Notes in Statistics, New York: Springer, 168.

Rivers, D. (2007). Sampling from web surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandria, VA.

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/en/pub/12-001-x/2000001/article/5179-eng.pdf.

Royall, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.

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

Särndal, C.-E., Lumiste, K. and Traat, I. (2016). Reducing the response imbalance: Is the accuracy of the survey estimates improved? *Survey Methodology*, 42, 2, 219-238. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2016002/article/14663-eng.pdf.

Schonlau, M., and Couper, M.P. (2017). Options for conducting Web surveys. *Statistical Science*, 32, 279-292.

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/en/pub/12-001-x/2013001/article/11824-eng.pdf.

Squire, P. (1988). Why the 1936 *Literary Digest* Poll failed. *Public Opinion Quarterly*, 52, 125-133.

Tourangeau, R., Brick, J.M., Lohr, S. and Li, J. (2017). Adaptive and responsive survey designs: A review and assessment. *Journal of the Royal Statistical Society*, 180, 201-223.

Valliant, R., and Dever, J.A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research*, 40, 105-137.

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

Wolter, K.M. (2007). *Introduction to Variance Estimation*. Second Edition, New-York: Springer.

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

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

Zhang, L.C. (2012). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*, 66, 41-63.

Vol. 46, No. 1, pp. 29-49
Statistics Canada, Catalogue No. 12-001-X

Local polynomial estimation for a small area mean under informative sampling

Marius Stefan and Michael A. Hidiroglou[1]

Abstract

Model-based methods are required to estimate small area parameters of interest, such as totals and means, when traditional direct estimation methods cannot provide adequate precision. Unit level and area level models are the most commonly used ones in practice. In the case of the unit level model, efficient model-based estimators can be obtained if the sample design is such that the sample and population models coincide: that is, the sampling design is non-informative for the model. If on the other hand, the sampling design is informative for the model, the selection probabilities will be related to the variable of interest, even after conditioning on the available auxiliary data. This will imply that the population model no longer holds for the sample. Pfeffermann and Sverchkov (2007) used the relationships between the population and sample distribution of the study variable to obtain approximately unbiased semi-parametric predictors of the area means under informative sampling schemes. Their procedure is valid for both sampled and non-sampled areas. Verret, Rao and Hidiroglou (2015) studied alternative procedures that incorporate a suitable function of the unit selection probabilities as an additional auxiliary variable. Their procedure resulted in approximately unbiased empirical best linear unbiased prediction (EBLUP) estimators for the small area means. In this paper, we extend the Verret et al. (2015) procedure by not assuming anything about the inclusion probabilities. Rather, we incorporate them into the unit level model via a smooth function of the inclusion probabilities. This function is estimated via a local approximation resulting in a local polynomial estimator. A conditional bootstrap method is proposed for the estimation of mean squared error (MSE) of the local polynomial and EBLUP estimators. The bias and efficiency properties of the local polynomial estimator are investigated via a simulation. Results for the bootstrap estimator of MSE are also presented.

Key Words: Local polynomial estimation; EBLUP estimation; Augmented model; Nested error model; Informative sampling; Conditional bootstrap.

1 Introduction

Population totals and means are often required for small subpopulations (or areas). When the inference is based on the area specific sample data, the resulting small area parameter estimators (direct estimators) are not of adequate precision due to the small area specific sample sizes. As a result, it becomes necessary to borrow strength across areas. Indirect estimators (predictors) that borrow strength are obtained when a model is used for the population of small areas. The model provides a link to related small areas. As a consequence, a model-based small area indirect estimator uses all the observations in the national sample, as well as the observations from the small area.

Suppose that the population of interest, U of size $N,$ consists of M non-overlapping areas with N_i units in the i^{th} small area $U_i \ (i = 1, \ldots, M)$. A sample, $s,$ of m areas is first selected using a specified sampling scheme with inclusion probabilities $\pi_i = mp_i \ (i = 1, \ldots, M),$ where p_i denotes the selection probability of small area $i.$ Subsamples s_i of specified sizes n_i are independently selected from each small area U_i according to a specified sampling design with selection probabilities $p_{j|i}$ $\left(\sum_{j=1}^{N_i} p_{j|i} = 1\right)$. The inclusion probabilities are $\pi_{j|i} = n_i p_{j|i}$ with sampling weights $w_{j|i} = \pi_{j|i}^{-1}.$ We consider the selection probabilities $p_{j|i}$ proportional to a size measure, $c_{ij},$ related to the response

[1] Marius Stefan, Faculty of Applied Sciences, Polytechnic University of Bucharest, Splaiul Independentei, nr. 313. E-mail: mastefan@gmail.com; Michael A. Hidiroglou, Statistics Canada Alumni. E-mail: hidirog@yahoo.ca.

variable y_{ij}: that is $p_{j|i} = c_{ij} \Big/ \sum_{k=1}^{N_i} c_{ik}$. We assume that all small areas are sampled, that is $m = M$. The resulting overall sample size is $n = \sum_{i=1}^{M} n_i$.

The basic population nested error regression model introduced by Battese, Harter and Fuller (1988) is given by

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

where y_{ij} is the value of the response variable for unit j in small area i, $\mathbf{x}_{ij} = \left(1, \; x_{ij1}, \ldots, x_{ijp}\right)^T$ is the vector of covariates, $\boldsymbol{\beta} = \left(\beta_0, \beta_1, \ldots, \beta_p\right)^T$ is the vector of fixed effects, and $v_i \overset{\text{iid}}{\sim} N\left(0, \sigma_v^2\right)$ are the random small area effects independent of the unit level errors $e_{ij} \overset{\text{iid}}{\sim} N\left(0, \sigma_e^2\right)$. The estimation of small area means, $\bar{Y}_i = N_i^{-1} \sum_{j=1}^{N_i} y_{ij}$, is of primary interest.

If the sampling design is non-informative for the model, that is if the model (1.1) holds for the sample, then efficient model-based estimators of the small area means $\bar{Y}_i$ can be obtained using empirical best linear unbiased prediction (EBLUP) (see Rao and Molina, 2015, Chapter 6 for an excellent account of the procedure). In this case, both the sample and population models coincide, allowing the use of (1.1) on the sample data to estimate $\bar{Y}_i$.

If the selection probability $p_{j|i}$ is related to y_{ij} even after conditioning on $\mathbf{x}_{ij}$, the sampling design is informative and the model (1.1) no longer holds for the sample. Consequently, the EBLUP estimator, that is based on (1.1) for the sample, may be heavily biased. It is, therefore, necessary to develop estimators that can account for sample selection, thereby reducing estimation bias. To this end, Verret et al. (2015) augmented model (1.1) by including the variable $g\left(p_{j|i}\right)$, where $g\left(p_{j|i}\right)$ is a specified function of the probability $p_{j|i}$. Their model for the sample is given by

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta}_0 + g\left(p_{j|i}\right) \delta_0 + v_{0i} + e_{0ij}, \; j = 1, \ldots, n_i; \; i = 1, \ldots, M, \tag{1.2}$$

where $v_{0i} \overset{\text{iid}}{\sim} N\left(0, \sigma_{0v}^2\right)$ and independent of $e_{0ij} \overset{\text{iid}}{\sim} N\left(0, \sigma_{0e}^2\right)$, and $\boldsymbol{\beta}_0 = \left(\beta_{00}, \beta_{01}, \ldots, \beta_{0p}\right)^T$. Verret et al. (2015) checked the adequacy of (1.2) after fitting the model to sample data $\left(y_{ij}, \mathbf{x}_{ij}, p_{j|i}\right)$, $j = 1, \ldots, n_i$; $i = 1, \ldots, M$, for different choices of $g(\cdot)$ that provide the best fit to the data. They suggested the following four possibilities for the choice of $g\left(p_{j|i}\right)$: $p_{j|i}$, $\log\left(p_{j|i}\right)$, $w_{j|i} = \left(n_i p_{j|i}\right)^{-1}$ and $n_i w_{j|i} = p_{j|i}^{-1}$. Since their sample model is parametric, the EBLUP theory can be used to estimate the relevant parameters using model (1.2).

Verret et al. (2015) illustrated via a simulation that the resulting EBLUP estimator, denoted as $\hat{\bar{Y}}_i^{\text{VRH}}$, obtained under (1.2), performs well under informative sampling design by reducing both bias and mean squared error as compared to the EBLUP estimator, $\hat{\bar{Y}}_i^{\text{EBLUP}}$, obtained from the sample data under the non-augmented model (1.1). Their simulation study compared their approach to the one used in Pfeffermann and Sverchkov (2007). Their simulation results showed that the bias-adjusted estimator of Pfeffermann and Sverchkov (2007) performed well under informative sampling in terms of bias, but that its MSE is significantly larger than the corresponding MSE of the EBLUP estimator based on the augmented model.

In this paper, we make no assumptions concerning the form of the function $g\left(p_{j|i}\right)$. Instead, we incorporate the $p_{j|i}$'s into the model (1.1) via an unknown smooth function $m_0\left(p_{j|i}\right)$. Our smooth function $m_0(\cdot)$ does not have a parametric form such as the one in Verret et al. (2015). We suppose that $m_0(\cdot)$ can be locally approximated by a polynomial of order q. For each point l in small area U_k, the corresponding polynomial is obtained by the Taylor expansion of $m_0\left(p_{j|i}\right)$ in a neighbourhood of $p_{l|k}$. For each point (l, k) in the population, we replace $m_0\left(p_{j|i}\right)$ by the corresponding parametric approximation and fit the resulting model just as in parametric fitting. We refer to this method as parametric polynomial localization.

This local approximation results in an augmented model that is semi-parametric. Such models have been applied to small area estimation by Opsomer, Claeskens, Ranalli, Kauermann and Breidt (2008). These authors chose a technique based on penalized splines to estimate the non-parametric part of their models. Breidt and Opsomer (2000) and Breidt, Opsomer, Johnson and Ranalli (2007) used the local polynomial technique in survey sampling theory to construct model-assisted estimators. Their estimators were based on non-parametric models without random effects. To the best of our knowledge, the estimation of a small area mean $\bar{Y}_i$, based on a local polynomial technique under semiparametric models has hardly been investigated.

The paper is structured as follows. Section 2 provides a review of two methods that result in estimators that account for sample selection: these methods were developed by Pfeffermann and Sverchkov (2007) and by Verret et al. (2015). In Section 3, we present a three-step procedure to estimate the proposed semi-parametric augmented model and the small area mean $\bar{Y}_i$ using a local polynomial approximation. We label the resulting estimator of the small area mean as $\hat{\bar{Y}}_i^{\text{LP}}$. The mean squared error (or MSE) of $\hat{\bar{Y}}_i^{\text{LP}}$ is estimated in Section 4 by a parametric conditional bootstrap method. The conditional bootstrap method is also used to estimate the MSE of EBLUP estimators obtained under augmented model (1.2). In Section 5, we conduct a simulation study under the design-model (or pm) framework to compare the bias and MSE of the new estimator $\hat{\bar{Y}}_i^{\text{LP}}$ to the EBLUP estimator, as well as to the two estimators discussed in Verret et al. (2015). We also study the performance of the conditional bootstrap procedure in estimating the MSE of the proposed local polynomial and EBLUP estimators studied in Verret et al. (2015). The performance is evaluated in terms of mean relative bias and mean confidence interval level. Concluding remarks are given in Section 6.

2 Existing methods

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

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

where $\hat{\gamma}_i = \hat{\sigma}_v^2 \big/ \left(\hat{\sigma}_v^2 + \hat{\sigma}_e^2 / n_i\right)$, $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij} \Big/ n_i$, $\bar{\mathbf{x}}_i = \sum_{j=1}^{n_i} \mathbf{x}_{ij} \Big/ n_i$ are the unweighted sample means of the response variable y and the covariates $\mathbf{x}$, and $\hat{v}_i = \hat{\gamma}_i \left(\bar{y}_i - \bar{\mathbf{x}}_i^T \hat{\boldsymbol{\beta}}\right)$. The estimator of the regression vector $\boldsymbol{\beta}$ in (1.1) is

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

The estimated variance components $\left(\hat{\sigma}_v^2, \hat{\sigma}_e^2\right)$ are obtained by the Henderson method of fitting of constants (HFC) or restricted maximum likelihood (REML) (see Battese et al., 1988 and Chapter 7 in Rao and Molina, 2015). The EBLUP estimator of the area mean $\bar{Y}_i$ may be written in terms of $\hat{\mu}_i^{\text{EBLUP}}$ as

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

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

Pfeffermann and Sverchkov (2007) studied the estimation of small area means under informative sampling, assuming the following model for the sample data

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\alpha} + u_i + h_{ij}; \; j = 1, \ldots, n_i; \; i = 1, \ldots, M, \tag{2.4}$$

where $u_i \overset{\text{iid}}{\sim} N\left(0, \sigma_u^2\right)$, and $h_{ij} \middle| j \in s_i \overset{\text{iid}}{\sim} N\left(0, \sigma_h^2\right)$. They assumed that the unit design weight $w_{j|i} = \pi_{j|i}^{-1}$ is random with conditional expectation

$$\begin{aligned} E_{si}\left(w_{j|i} \middle| \mathbf{x}_{ij}, y_{ij}, v_i \right) &= E_{si}\left(w_{j|i} \middle| \mathbf{x}_{ij}, y_{ij} \right) \\ &= k_i \exp\left(\mathbf{x}_{ij}^T \mathbf{a} + d y_{ij} \right), \end{aligned} \tag{2.5}$$

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

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

The Pfeffermann and Sverchkov (2007) estimator of $\bar{Y}_i$ provides protection against informative sampling supposing that this assumption holds. The estimator is given by

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

where $\hat{\mu}_{iu}^{\text{EBLUP}} = \bar{\mathbf{X}}_i^T \hat{\boldsymbol{\alpha}} + \hat{u}_i$ is the EBLUP estimator of $\mu_{iu} = \bar{\mathbf{X}}_i^T \boldsymbol{\alpha} + u_i$ under the sample model (2.4) and $\hat{d}$ is an estimator of d in the model (2.5) for the weights $w_{j|i}$. The last term in (2.6) corrects for any bias due to informative sampling under (2.5). Pfeffermann and Sverchkov (2007) obtained the estimator $\hat{d}$ of d in (2.5) by regressing the sampling weights $w_{j|i}$ on $k_i \exp\left(\mathbf{x}_{ij}^T \mathbf{a} + d y_{ij} \right)$. The coefficients k_i, $\mathbf{a}$ and d may be estimated by fitting the model (2.5) using the NLIN procedure in SAS or function `nls` in Splus. This involves iterative calculations and the initial values for $\mathbf{a}$ and d are obtained by regressing $\log\left(w_{j|i} \right)$ on $\mathbf{x}_{ij}$ and y_{ij}. Initial values for $\hat{k}_i$, $i = 1, \ldots, M$ are taken as $k_i = N_i / n_i$.

The Verret et al. (2015) estimator is obtained when the EBLUP theory is applied to model (1.2). Let $\mathbf{x}_{ij}^{\text{aug}} = \left(\mathbf{x}_{ij}^T, g\left(p_{j|i}\right)\right)^T$ be the vector $\mathbf{x}_{ij}$ augmented by the variable $g\left(p_{j|i}\right)$, $\bar{G}_i$ the area mean of the population values $g\left(p_{j|i}\right)$, and $\mu_{0i} = \bar{\mathbf{X}}_i^T\boldsymbol{\beta}_0 + \bar{G}_i\delta_0 + v_{0i}$. The EBLUP estimator of μ_{0i} is given by

$$\hat{\mu}_{0i}^{\text{EBLUP}} = \bar{\mathbf{X}}_i^T\hat{\boldsymbol{\beta}}_0 + \bar{G}_i\hat{\delta}_0 + \hat{v}_{0i} = \hat{\gamma}_{0i}\bar{y}_i + \left(\bar{\mathbf{X}}_i - \hat{\gamma}_{0i}\bar{\mathbf{x}}_i\right)^T\hat{\boldsymbol{\beta}}_0 + \left(\bar{G}_i - \hat{\gamma}_{0i}\bar{g}_i\right)\hat{\delta}_0, \tag{2.7}$$

where $\hat{\gamma}_{0i} = \hat{\sigma}_{0v}^2 / \left(\hat{\sigma}_{0v}^2 + \hat{\sigma}_{0e}^2 / n_i\right)$, $\bar{g}_i = \sum_{j=1}^{n_i} g\left(p_{j|i}\right) / n_i$ and $\hat{v}_{0i} = \hat{\gamma}_{0i}\left(\bar{y}_i - \bar{\mathbf{x}}_i^T\hat{\boldsymbol{\beta}}_0 - \bar{g}_i\hat{\delta}_0\right)$. The parameters, $(\boldsymbol{\beta}_0, \delta_0)$ are estimated by

$$\left(\hat{\boldsymbol{\beta}}_0^T, \hat{\delta}_0\right)^T = \left\{\sum_{i=1}^{M}\sum_{j=1}^{n_i}\mathbf{x}_{ij}^{\text{aug}}\left(\mathbf{x}_{ij}^{\text{aug}} - \hat{\gamma}_{0i}\bar{\mathbf{x}}_i^{\text{aug}}\right)^T\right\}^{-1}\left\{\sum_{i=1}^{M}\sum_{j=1}^{n_i}\left(\mathbf{x}_{ij}^{\text{aug}} - \hat{\gamma}_{0i}\bar{\mathbf{x}}_i^{\text{aug}}\right)y_{ij}\right\}, \tag{2.8}$$

with $\bar{\mathbf{x}}_i^{\text{aug}} = \sum_{j=1}^{n_i}\mathbf{x}_{ij}^{\text{aug}} \big/ n_i = \left(\bar{\mathbf{x}}_i^T, \bar{g}_i\right)^T$. The model parameters $\left(\hat{\sigma}_{0v}^2, \hat{\sigma}_{0e}^2\right)$ are estimated by HFC or REML method. The estimator of the area mean $\bar{Y}_i$, denoted $\hat{\bar{Y}}_i^{\text{VRH}}$, may be written in terms of $\hat{\mu}_{0i}^{\text{EBLUP}}$ as

$$\hat{\bar{Y}}_i^{\text{VRH}} = \frac{1}{N_i}\left[\left(N_i - n_i\right)\hat{\mu}_{0i}^{\text{EBLUP}} + n_i\left\{\bar{y}_i + \left(\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_i\right)^T\hat{\boldsymbol{\beta}}_0 + \left(\bar{G}_i - \bar{g}_i\right)^T\hat{\delta}_0\right\}\right]. \tag{2.9}$$

3 The local polynomial estimator

3.1 The estimation of a small area mean

The objective is to estimate the mean $\bar{Y}_i$ for small area U_i for $i = 1, \ldots, M$. Splitting the population U_i into observed units in the sample, s_i of size n_i, and non-observed units in the non-sampled portion, $\bar{s}_i = U_i / s_i$ of size $N_i - n_i$, we can express $\bar{Y}_i$ as

$$\bar{Y}_i = \frac{1}{N_i}\left(\sum_{j\in s_i} y_{ij} + \sum_{j\in\bar{s}_i} y_{ij}\right). \tag{3.1}$$

Given that we do not know the y values for the non-observed units in sets $\bar{s}_i$ for $i = 1, \ldots, M$, we need to estimate them. Denoting as $\hat{y}_{ij}$ the estimator of y_{ij} for such units, the resulting estimator of the mean $\bar{Y}_i$ is

$$\hat{\bar{Y}}_i = \frac{1}{N_i}\left(\sum_{j\in s_i} y_{ij} + \sum_{j\in\bar{s}_i}\hat{y}_{ij}\right). \tag{3.2}$$

We obtain estimators $\hat{y}_{ij}$ of y_{ij}, for $j \in \bar{s}_i$, based on an augmented model that includes an unknown smooth function of the selection probabilities $p_{j|i}$, denoted $m_0\left(p_{j|i}\right)$. The proposed augmented semi-parametric sample model is given by

$$y_{ij} = \tilde{\mathbf{x}}_{ij}^T\boldsymbol{\beta}_1 + m_0\left(p_{j|i}\right) + v_{1i} + e_{1ij},\ j = 1, \ldots, n_i;\ i = 1, \ldots, M, \tag{3.3}$$

where $v_{1i} \overset{\text{iid}}{\sim} N(0, \sigma_{1v}^2)$ and independent of $e_{1ij} \overset{\text{iid}}{\sim} N(0, \sigma_{1e}^2)$. The vector $\tilde{\mathbf{x}}_{ij} = (x_{ij1}, \ldots, x_{ijp})^T$ in model (3.3) represents the covariates $\mathbf{x}_{ij}$ without a constant (i.e., the intercept) and $\boldsymbol{\beta}_1 = (\beta_{11}, \ldots, \beta_{1p})^T$ a vector of fixed effects. Model (3.3) is semi-parametric as the response variable y_{ij} depends linearly on the vector of auxiliary variables, $\tilde{\mathbf{x}}_{ij}$, and the probability of selection $p_{j|i}$ enters non-parametrically through the smooth function $m_0(\cdot)$.

We assume that model (3.3) has a similar covariance structure with the one associated with model (1.2): the small area effects v_{1i} and random errors e_{1ij} are iid, normally distributed and independently of one another. However, the semi-parametric model (3.3) is more flexible than the parametric model (1.2), as it does not force the function $m_0(p_{j|i})$ to be of a specific form. There is a disadvantage to this set-up. Since model (3.3) is not a linear mixed model, the general EBLUP theory given in Section 2 cannot be applied directly to obtain estimators of $m_0(p_{j|i})$, $\boldsymbol{\beta}_1$ and v_{1i}. Consequently, we propose to estimate (3.3) by combining the EBLUP theory for linear mixed models and the local polynomial technique (Fan and Gijbels, 1996).

We estimate (3.3) in three steps. In the first step, we obtain estimates of $m_0(p_{j|i})$, $\hat{m}_0(p_{j|i})$, $j = 1, \ldots, N_i$, $i = 1, \ldots, M$, for all units in the population. These estimates are local in character as they are based on the local polynomial technique. Estimates $\hat{m}_0(p_{j|i})$, $j \in s_i$ for the observed units are then used in the second step to obtain global estimators of $\boldsymbol{\beta}_1$ and v_{1i}, $i = 1, \ldots, M$. We denote these estimators as $\hat{\boldsymbol{\beta}}_{\text{glo},1}$ and $\hat{v}_{\text{glo},1i}$, $i = 1, \ldots, M$. Finally, in the third step, we use the local estimators $\hat{m}_0(p_{j|i})$ for the unobserved units, obtained in the first step, and the global estimators $\hat{\boldsymbol{\beta}}_{\text{glo},1}$ and $\hat{v}_{\text{glo},1i}$ obtained in the second step, to estimate y_{ij} for $j \in \bar{s}_i$ and $i = 1, \ldots, M$. The resulting estimators of y_{ij}, denoted as $\hat{y}_{ij}$, are

$$\hat{y}_{ij} = \tilde{\mathbf{x}}_{ij}^T \hat{\boldsymbol{\beta}}_{\text{glo},1} + \hat{m}_0(p_{j|i}) + \hat{v}_{\text{glo},1i}, \ j \in \bar{s}_i. \tag{3.4}$$

The $\hat{y}_{ij}$'s are incorporated into equation (3.2) to obtain the estimator of the small area mean $\hat{\bar{Y}}_i$.

We now proceed to describe the first step in more detail. Following Ruppert and Matteson (2015), we estimate the values of the unknown function $m_0(p_{l|k})$ for all units $l \in U_k$ and small areas k, with $k = 1, \ldots, M$, by using local polynomial regression. Local polynomial regression is based on the principle that a smooth function can be approximated locally by a low-degree polynomial. We approximate $m_0(p_{j|i})$ in model (3.3) by a q^{th}-degree polynomial, say $m_1(p_{j|i})$, using a Taylor expansion around $p_{l|k}$. The approximation is given by

$$m_1(p_{j|i}) = m_0(p_{l|k}) + \sum_{a=1}^{q} \frac{1}{a!} m_0(p_{l|k})^{(a)} (p_{j|i} - p_{l|k})^a, \ j \in s_i; \ i = 1, \ldots, M, \tag{3.5}$$

where $m_0(p_{l|k})^{(a)}$ is the a^{th} derivative of $m_0(p_{j|i})$ evaluated at $p_{l|k}$. The function $m_1(p_{j|i})$ depends on $l \in U_k$, but we suppress this dependence to simplify the notation.

For each point $p_{l|k}$, $l \in U_k$; $k = 1, \ldots, M$, in model (3.3) we replace $m_0(p_{j|i})$ by its approximation $m_1(p_{j|i})$ given by (3.5). The resulting model is given by

$$y_{ij} \;=\; \tilde{\mathbf{x}}_{ij}^{T}\boldsymbol{\beta}_1 + m_0\left(p_{l|k}\right) + \sum_{a=1}^{q}\frac{1}{a!}m_0\left(p_{l|k}\right)^{(a)}\left(p_{j|i} - p_{l|k}\right)^{a} + v_{1i} + e_{1ij},\;\; j \in s_i;\; i = 1, \ldots, M. \tag{3.6}$$

Model (3.6) is an approximate local model for (3.3) depending on the point $l \in U_k$ of the population. Estimates of $\boldsymbol{\beta}_1$ and v_{1i} based on (3.6) will be denoted by $\hat{\boldsymbol{\beta}}_{\text{loc},1}$ and $\hat{v}_{\text{loc},1i}$. Notice that (3.6) allows the estimation of $m_0\left(p_{l|k}\right)$, the value of the smooth function $m_0(\cdot)$ at a point $p_{l|k}$. We express (3.6) as

$$y_{ij} \;=\; \tilde{\mathbf{x}}_{ij}^{T}\boldsymbol{\beta}_1 \;+\; u_0 \;+\; \sum_{a=1}^{q}u_a\left(p_{j|i} \;-\; p_{l|k}\right)^{a} \;+\; v_{1i} \;+\; e_{1ij}\text{:}\; j \;\in\; s_i;\; i \;=\; 1, \ldots, M, \tag{3.7}$$

where $u_a \;=\; m_0\left(p_{l|k}\right)^{(a)}\big/\,a!$ for $a \;=\; 0, \ldots, q$. Model (3.7) is a linear mixed model with fixed parameters $\left(\boldsymbol{\beta}_1,\; u_0, \ldots, u_q\right)$ and random small area effects $v_{1i},\; i \;=\; 1, \ldots, M$.

Let $\hat{u}_0$ be an estimator of u_0 obtained by fitting model (3.7). An approximate estimator of $m_0\left(p_{l|k}\right) \;=\; u_0$ is given by $\hat{m}_0\left(p_{l|k}\right) \;=\; \hat{u}_0$. Since we require estimators of $m_0\left(p_{l|k}\right)$ for $l \in U_k$ and $k \;=\; 1, \ldots, M$, we use $N \;=\; \sum_{i=1}^{M}N_i$ models (3.7). As pointed out by an Associate Editor, if N is large, estimating the values of $m_0(\cdot)$ for all points in the population can be computationally intensive.

It is more convenient to work with matrix notation. To this end, we define $\mathbf{y}_i \;=\; \left(y_{i1}, \ldots, y_{in_i}\right)^T$, $\tilde{\mathbf{X}}_i = \left(\tilde{\mathbf{x}}_{i1}^{T}, \ldots, \tilde{\mathbf{x}}_{in_i}^{T}\right)^T$, $\mathbf{m}_{0,i} = \left(m_0\left(p_{1|i}\right), \ldots, m_0\left(p_{n_i|i}\right)\right)^T$, $\mathbf{v}_1 = \left(v_{11}, \ldots, v_{1M}\right)^T$ and $\mathbf{e}_{1i} = \left(e_{1i1}, \ldots, e_{1in_i}\right)^T$. Model (3.3) can be expressed in a matrix form by stacking the observations, and the resulting equation is

$$\mathbf{y} \;=\; \tilde{\mathbf{X}}\boldsymbol{\beta}_1 \;+\; \mathbf{m}_0 \;+\; \mathbf{Z}\mathbf{v}_1 \;+\; \mathbf{e}_1, \tag{3.8}$$

where $\mathbf{y} \;=\; \text{col}_{1\le i\le M}\left(\mathbf{y}_i\right)$, $\tilde{\mathbf{X}} \;=\; \text{col}_{1\le i\le M}\left(\tilde{\mathbf{X}}_i\right)$, $\mathbf{m}_0 \;=\; \text{col}_{1\le i\le M}\left(\mathbf{m}_{0,i}\right)$, $\mathbf{Z} \;=\; \text{diag}_{1\le i\le M}\left\{\mathbf{1}_{n_i}\right\}$ and $\mathbf{e}_1 \;=\; \text{col}_{1\le i\le M}\left(\mathbf{e}_{1i}\right)$.

For unit l in small area U_k, we define the $n \times (q + 1)$ matrix:

$$\mathbf{Q} \;=\; \begin{pmatrix} 1 & \left(p_{1|1} - p_{l|k}\right) & \cdots & \left(p_{1|1} - p_{l|k}\right)^q \\ \vdots & \vdots & \cdots & \vdots \\ 1 & \left(p_{n_M|M} - p_{l|k}\right) & \cdots & \left(p_{n_M|M} - p_{l|k}\right)^q \end{pmatrix},$$

where $n \;=\; \sum_{i=1}^{M}n_i$ is the total sample size. Let $\mathbf{u} \;=\; \left(m_0\left(p_{l|k}\right),\, m_0^{(1)}\left(p_{l|k}\right)/1!, \ldots, m_0^{(q)}\left(p_{l|k}\right)/q!\right)^T$ represent the vector of derivatives of the function $m_0(\cdot)$ evaluated at $p_{l|k}$. The terms $\mathbf{Q}$ and $\mathbf{u}$ depend on the unit $l \in U_k$ where the localization is realized. We omitted their dependence on the unit l from small area U_k in order not to burden the notation. We define vector $\mathbf{m}_1$ obtained by stacking the n values of the function $m_1(\cdot)$ defined by (3.5). That is, $\mathbf{m}_1 \;=\; \text{col}_{1\le i\le M}(\mathbf{m}_{1,i})$ with $\mathbf{m}_{1,i} \;=\; (m_1(p_{1|i}), \ldots, m_1(p_{n_i|i}))^T$. This allows to approximate $\mathbf{m}_0$ by $\mathbf{m}_0 \;\approx\; \mathbf{m}_1$. The vector $\mathbf{m}_1$ is given by $\mathbf{m}_1 \;=\; \mathbf{Qu}$. It then follows that an approximation to (3.8) in a neighbourhood of $l \in U_k$ is

$$\mathbf{y} \;=\; \tilde{\mathbf{X}}\boldsymbol{\beta}_1 \;+\; \mathbf{Qu} \;+\; \mathbf{Z}\mathbf{v}_1 \;+\; \mathbf{e}_1. \tag{3.9}$$

Equations (3.8) and (3.9) are the matrix form equivalents of equations (3.3) and (3.7), respectively. The matrix $\tilde{\mathbf{X}}$ in (3.9) does not include the constant term that represents the intercept, because this term is already included in $\mathbf{Q}$. Equation (3.9) is a standard linear mixed effects model with fixed parameters $\boldsymbol{\beta}_{\text{fixed}} = (\boldsymbol{\beta}_1^T, \mathbf{u}^T)^T$ and random small area effects $\mathbf{v}_1$. We denote by $V(\mathbf{v}_1) = \mathbf{G} = \sigma_{1v}^2 \mathbf{I}_M$, $V(\mathbf{e}_{1i}) = \mathbf{R}_i = \sigma_{1e}^2 \mathbf{I}_{n_i}$ and $V(\mathbf{e}_1) = \mathbf{R} = \text{diag}_{1 \le i \le M} \{\mathbf{R}_i\}$ as the respective covariance matrices of $\mathbf{v}_1$, $\mathbf{e}_{1i}$ and $\mathbf{e}_1$. The covariance matrix of $\mathbf{y}_i$ is given by $V(\mathbf{y}_i) = \mathbf{V}_i = \sigma_{1v}^2 \mathbf{J}_{n_i} + \sigma_{1e}^2 \mathbf{I}_{n_i}$. The matrices $\mathbf{I}_M$ and $\mathbf{I}_{n_i}$ are the identity matrices of order M and n_i respectively, whereas $\mathbf{J}_{n_i}$ is the square matrix of order n_i with all its elements equal to 1. It follows that $V(\mathbf{y}) = \mathbf{V} = \text{diag}_{1 \le i \le M} \{\mathbf{V}_i\}$.

Assume that $\mathbf{V}$ is known and that $\mathbf{v}_1$ and $\mathbf{e}_1$ are normally distributed. Using classical EBLUP theory, estimators of $\boldsymbol{\beta}_{\text{fixed}}$ and $\mathbf{v}_1$ can be obtained by minimizing

$$\Phi = (\mathbf{y} - \tilde{\mathbf{X}}\boldsymbol{\beta}_1 - \mathbf{Qu} - \mathbf{Zv}_1)^T \mathbf{R}^{-1} (\mathbf{y} - \tilde{\mathbf{X}}\boldsymbol{\beta}_1 - \mathbf{Qu} - \mathbf{Zv}_1) + \mathbf{v}_1^T \mathbf{G}^{-1} \mathbf{v}_1.$$

Note that all the observations that are included in Φ are equally weighted. However, we need to modify Φ to be in line with how local polynomial estimation is carried out. To this end, referring back to equation (3.7), we estimate its parameters by associating kernel weights $K((p_{j|i} - p_{l|k})/h)/h$ to each sampled unit $j \in s_i$; $i = 1, \dots, M$. These kernel weights are chosen so as to give a larger weight to the sample points that are close to $l \in U_k$, and a smaller weight to those that are further away. The weight $K(\cdot)$ is a probability density function and h is a bandwidth controlling the size of the local neighbourhood. We explain in Section 3.2 how an optimal bandwidth can be obtained. Let $\mathbf{W}$ be the $n \times n$ diagonal matrix of kernel weights given by

$$\mathbf{W} = \underset{\substack{1 \le j \le n_i \\ 1 \le i \le M}}{\text{diag}} \left\{ \frac{1}{h} K\left(\frac{p_{j|i} - p_{l|k}}{h} \right) \right\}.$$

The matrix $\mathbf{W}$ depends on unit l from small area U_k and the bandwidth h. We do not include the subscripts $l \in U_k$ and h in the definition of the matrix $\mathbf{W}$, in order not to burden notation. Following Wu and Zhang (2002), the incorporation of the kernel weights in Φ lead us to minimize Φ_W where

$$\Phi_W = (\mathbf{y} - \tilde{\mathbf{X}}\boldsymbol{\beta}_1 - \mathbf{Qu} - \mathbf{Zv}_1)^T \mathbf{W}^{1/2} \mathbf{R}^{-1} \mathbf{W}^{1/2} (\mathbf{y} - \tilde{\mathbf{X}}\boldsymbol{\beta}_1 - \mathbf{Qu} - \mathbf{Zv}_1) + \mathbf{v}_1^T \mathbf{G}^{-1} \mathbf{v}_1,$$

and $\mathbf{W}^{1/2}$ represents the square root of the matrix $\mathbf{W}$.

Estimating the parameters of (3.9) by minimizing Φ_W is equivalent to estimating those given by

$$\mathbf{W}^{1/2}\mathbf{y} = \mathbf{W}^{1/2}\tilde{\mathbf{X}}\boldsymbol{\beta}_1 + \mathbf{W}^{1/2}\mathbf{Qu} + \mathbf{W}^{1/2}\mathbf{Zv}_1 + \mathbf{e}_1. \tag{3.10}$$

The weighted EBLUP based on (3.9) with the matrix of weights given by $\mathbf{W}$ corresponds to a classical EBLUP obtained from model (3.10). Define $\mathbf{y}_w = \mathbf{W}^{1/2}\mathbf{y}$, $\mathbf{X}_w = [\mathbf{W}^{1/2}\tilde{\mathbf{X}}, \mathbf{W}^{1/2}\mathbf{Q}]$ and $\mathbf{Z}_w = \mathbf{W}^{1/2}\mathbf{Z}$. Equation (3.10) can be rewritten as

$$\mathbf{y}_w = \mathbf{X}_w \boldsymbol{\beta}_{\text{fixed}} + \mathbf{Z}_w \mathbf{v}_1 + \mathbf{e}_1. \tag{3.11}$$

Let $\hat{\boldsymbol{\beta}}_{\text{loc, fixed}} = (\hat{\boldsymbol{\beta}}_{\text{loc},1}^T, \hat{\mathbf{u}}^T)^T$ and $\hat{\mathbf{v}}_{\text{loc},1} = (\hat{v}_{\text{loc},11}, \dots, \hat{v}_{\text{loc},1M})^T$ be the EBLUP estimators of the fixed and random effects of (3.11). The estimators $\hat{\boldsymbol{\beta}}_{\text{loc, fixed}}$ and $\hat{\mathbf{v}}_{\text{loc},1}$ are based on local estimators of the variance components $(\sigma_{1v}^2, \sigma_{1e}^2)$. The estimators of these components, denoted as $(\hat{\sigma}_{\text{loc},1v}^2, \hat{\sigma}_{\text{loc},1e}^2)$, are obtained using HFC or REML methods under model (3.11). Given that $\mathbf{u} = (m_0(p_{l|k}), m_0^{(1)}(p_{l|k})/1!, \ldots, m_0^{(q)}(p_{l|k})/q!)^T$, an estimator $\hat{m}_0(p_{l|k})$ of $m_0(p_{l|k})$ is the first component $\hat{u}_0$ of $\hat{\mathbf{u}}$.

Notice that $\hat{\boldsymbol{\beta}}_{\text{loc},1}$, $\hat{m}_0(p_{l|k})$ and $\hat{v}_{\text{loc},1k}$ could be used to obtain local estimates $\hat{y}_{\text{loc},kl}$ for the unknown value y_{kl}, where $\hat{y}_{\text{loc},kl} = \tilde{\mathbf{x}}_{kl}^T \hat{\boldsymbol{\beta}}_{\text{loc},1} + \hat{m}_0(p_{l|k}) + \hat{v}_{\text{loc},1k}$ for $l \in \bar{s}_k$. However, a referee pointed out that, in practice, this methodology would not likely to be well behaved because it requires a strong balance of the small areas across the range of the probabilities $p_{l|k}$. If this balance is not respected, the resulting estimation would suffer severely from this localization. As a consequence, we opted for a global estimation of $\boldsymbol{\beta}_1$ and $\mathbf{v}_1$.

We now explain the second step of our procedure. Parameters $\boldsymbol{\beta}_1$ and $\mathbf{v}_1$ can be estimated globally based on the estimations $\hat{m}_0(p_{j|i})$ and the auxiliary data $\tilde{\mathbf{x}}_{ij}$ associated with the sample units. For $j \in s_i$ and $i = 1, \ldots, M$, define a new variable, say ξ, as

$$\xi_{ij} = y_{ij} - \hat{m}_0(p_{j|i}), \; j \in s_i; \; i = 1, \ldots, M.$$

The n values ξ_{ij} represent the differences between the observed y_{ij}'s and their local estimators $\hat{m}_0(p_{j|i})$. Using model (3.3), ξ satisfies the following model

$$\xi_{ij} = \tilde{\mathbf{x}}_{ij}^T \boldsymbol{\beta}_{\text{glo},1} + v_{\text{glo},1i} + e_{\text{glo},1ij}, \; j \in s_i; \; i = 1, \ldots, M, \tag{3.12}$$

where $v_{\text{glo},1i} \sim N(0, \sigma^2_{\text{glo},1v})$ and $e_{\text{glo},1ij} \sim N(0, \sigma^2_{\text{glo},1e})$. The subscript glo indicates that (3.12) is a global model.

Given that (3.12) represents a parametric linear mixed effects model, we can use the classical (unweighted) EBLUP to estimate its parameters. Let $\hat{\boldsymbol{\beta}}_{\text{glo},1}$ and $\hat{v}_{\text{glo},1i}$ be the respective empirical best linear unbiased estimators of $\boldsymbol{\beta}_{\text{glo},1}$ and $v_{\text{glo},1i}$. Let $(\hat{\sigma}^2_{\text{glo},1v}, \hat{\sigma}^2_{\text{glo},1e})$ be the estimators of the variance components $(\sigma^2_{\text{glo},1v}, \sigma^2_{\text{glo},1e})$ where HFC or REML can be used to estimate these parameters. We estimate $(\boldsymbol{\beta}_1, v_{1i}, \sigma^2_{1v}, \sigma^2_{1e})$ of model (3.3) by $(\hat{\boldsymbol{\beta}}_{\text{glo},1}, \hat{v}_{\text{glo},1i}, \hat{\sigma}^2_{\text{glo},1v}, \hat{\sigma}^2_{\text{glo},1e})$ using model (3.12). The global estimators $\hat{\boldsymbol{\beta}}_{\text{glo},1}$, $\hat{v}_{\text{glo},1i}$ and $(\hat{\sigma}^2_{\text{glo},1v}, \hat{\sigma}^2_{\text{glo},1e})$ are free of bias caused by informative sampling design because ξ_{ij} is no longer related to the $p_{j|i}$'s after conditioning on $\mathbf{x}_{ij}$.

The third step estimates the non observed y_{ij} values, for $j \in \bar{s}_i$ and $i = 1, \ldots, M$, by plugging into equation (3.4): i. the local estimators $\hat{m}_0(p_{j|i})$ for $j \in \bar{s}_i$, obtained in the first step, and ii. the global estimators $\hat{\boldsymbol{\beta}}_{\text{glo},1}$ and $\hat{v}_{\text{glo},1i}$ obtained in the second step. The resulting $\hat{y}_{ij}$'s, for $j \in \bar{s}_i$, are inserted into (3.2) to compute the estimator $\hat{\bar{Y}}_i$. Note that $\hat{\bar{Y}}_i$ requires $\tilde{\mathbf{x}}_{ij}$ and $p_{j|i}$ are known for all the units of the population. A referee pointed out that, in practice, this assumption may limit the applicability of the proposed procedure. This could be remedied if National Statistical Offices provided access to the selection probabilities of all units, as they may be needed in applications such as this one.

3.2 Bandwidth selection

Local polynomials require the specification of the kernel $K(\cdot)$, the order of the polynomial fit q, as well as the bandwidth h. Fan and Gijbels (1996) state that values of q larger than 1 do not bring a significant improvement as compared to the linear fit $(q = 1)$. Fan and Gijbels (1996) also state that the choice of h is far more important than the degree of the polynomial. In what follows, we use a normal density kernel, and chose q equal to one, as this leads to satisfactory results for most applications.

The optimal h is determined using the cross-validation method (CV). For a given h, compute the estimator of y_{ij} given by (3.4) using the sample that remains after the j^{th} unit has been removed from s_i. Denoting the resulting estimator of y_{ij} as $\tilde{y}_{ij}$, we follow Wu and Zhang (2002) and define the CV criterion as

$$\mathrm{CV}(h) = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{n_i}\sum_{j\in s_i}\left(y_{ij} - \tilde{y}_{ij}\right)^2.$$

The term $1/n_i$ takes into account the number of observations within small area U_i. The optimal bandwidth h_{opt} is obtained by minimizing the $\mathrm{CV}(h)$. Given h_{opt}, the local polynomial estimator of the small area mean $\bar{Y}_i$ given by (3.2) is denoted as $\hat{\bar{Y}}_i^{\text{LP}}$.

4 MSE estimation based on the bootstrap

The MSE estimation of small area estimators is a challenging problem even in the case of classical EBLUP estimators. The general EBLUP theory provides a closed form approximation to $\mathrm{MSE}\left(\hat{\bar{Y}}_i^{\text{EBLUP}}\right)$ based on a linearization method. Using this approximation, an estimator for $\mathrm{MSE}\left(\hat{\bar{Y}}_i^{\text{EBLUP}}\right)$ can be obtained (see Prasad and Rao, 1990 for details). Verret et al. (2015) used the closed form approximation to estimate the mean squared error estimator for $\hat{\bar{Y}}_i^{\text{VRH}}$ given in (2.9). This was possible because estimator $\hat{\bar{Y}}_i^{\text{VRH}}$ is a standard EBLUP obtained under a linear mixed model that includes the additional known variable $g\left(p_{j|i}\right)$. No new theory is needed to estimate the MSE of $\hat{\bar{Y}}_i^{\text{VRH}}$. In our case, given the repeated local estimation of model (3.6), it is not possible to obtain a closed-form approximation to the mean squared error of $\hat{\bar{Y}}_i^{\text{LP}}$, $\mathrm{MSE}\left(\hat{\bar{Y}}_i^{\text{LP}}\right)$, nor for its estimator $\mathrm{mse}\left(\hat{\bar{Y}}_i^{\text{LP}}\right)$. We used two variants of the bootstrap procedure to estimate the MSE of the small area estimators that we have discussed so far. For estimating the MSE of $\hat{\bar{Y}}_i^{\text{EBLUP}}$, we used an *unconditional* bootstrap, whereas for $\hat{\bar{Y}}_i^{\text{LP}}$, $\hat{\bar{Y}}_i^{\text{VRH1}}$ and $\hat{\bar{Y}}_i^{\text{VRH2}}$, we used a *conditional* bootstrap. We proceed to describe how each bootstrap type is computed.

We first describe the unconditional bootstrap. This is a variant of the parametric bootstrap of Hall and Maiti (2006), proposed by González-Manteiga, Lombardia, Molina, Morales and Santamaria (2008). This procedure can be used for estimating the MSE of $\hat{\bar{Y}}_i^{\text{EBLUP}}$ that is based on model (1.1) because the estimates of the various parameters in model (1.1) do not depend on the selection probabilities $p_{j|i}$: $j \in s_i$; $i = 1, \ldots, M$. The y values are predicted by generating $v_i^* \sim N\left(0, \hat{\sigma}_v^2\right)$ and $e_{ij}^* \sim N\left(0, \hat{\sigma}_e^2\right)$, where $\left(\hat{\sigma}_v^2, \hat{\sigma}_e^2\right)$ are the HFC or REML estimators of $\left(\sigma_v^2, \sigma_e^2\right)$. Using the EBLUP estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, bootstrap values of y_{ij} are obtained as

$$y_{ij}^* = \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}} + v_i^* + e_{ij}^*,\ j \in U_i;\ i = 1, \ldots, M. \tag{4.1}$$

The bootstrap version of the target parameter $\bar{Y}_i$ is computed as $\bar{Y}_i^* = N_i^{-1}\sum_{j=1}^{N_i} y_{ij}^*$. The bootstrap version of the EBLUP estimator $\hat{\bar{Y}}_i^{\text{EBLUP}}$ is given by

$$\hat{\bar{Y}}_i^{\text{EBLUP*}} = \frac{1}{N_i}\left(\sum_{j\in s_i} y_{ij}^* + \sum_{j\in\bar{s}_i}\hat{y}_{ij}^*\right),$$

where $\hat{y}_{ij}^* = \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}^* + \hat{v}_i^*$ and $\left(\hat{\boldsymbol{\beta}}^*, \hat{v}_i^*\right)$ are the EBLUP estimators of $(\boldsymbol{\beta}, v_i)$ that are based on $\left(y_{ij}^*, \mathbf{x}_{ij}\right)$, $j \in s_i$, for $i = 1, \ldots, M$. Repeating the above procedure B times, the bootstrap estimator of $\text{MSE}\left(\hat{\bar{Y}}_i^{\text{EBLUP}}\right)$ is

$$\text{mse}_{\text{boot}}\left(\hat{\bar{Y}}_i^{\text{EBLUP}}\right) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\bar{Y}}_i^{\text{EBLUP*}}(b) - \bar{Y}_i^*(b)\right)^2, \tag{4.2}$$

where $\hat{\bar{Y}}_i^{\text{EBLUP*}}(b)$ and $\bar{Y}_i^*(b)$ are the values of $\hat{\bar{Y}}_i^{\text{EBLUP*}}$ and $\bar{Y}_i^*$ for the b^{th} bootstrap replicate. Since the estimators $\left(\hat{\boldsymbol{\beta}}, \hat{\sigma}_v^2, \hat{\sigma}_e^2\right)$ are severely biased due to the informative sampling design, we expect that $\text{mse}_{\text{boot}}\left(\hat{\bar{Y}}_i^{\text{EBLUP}}\right)$ will be a biased estimator of $\text{MSE}\left(\hat{\bar{Y}}_i^{\text{EBLUP}}\right)$. This is because it is based on the population model (1.1), and that this model does not hold for the sample.

We now turn to the estimation of $\text{MSE}\left(\hat{\bar{Y}}_i^{\text{LP}}\right)$ via the conditional bootstrap. Recall that $\hat{\bar{Y}}_i^{\text{LP}}$ is based on the augmented model (3.3). It is therefore natural to use this model when we estimate the precision of the local polynomial estimator. It is not possible to use the parametric unconditional bootstrap as it would require the generation of bootstrap values $\left(y_{ij}^*, p_{j|i}^*\right)$ for both y_{ij} and $p_{j|i}$, and this would imply that we would need to know how the y_{ij}'s are related to the selection probabilities $p_{j|i}$. As the Associate Editor pointed out, the exact relationship between y_{ij} and $p_{j|i}$ is not known in practice. We therefore opted to keep the selection probabilities $p_{j|i}$ associated with the initial sample, and generate bootstrap values only for the response variable y_{ij}. The resulting bootstrap is conditional on $p_{j|i}$, $j \in U_i$; $i = 1, \ldots, M$, and it is for this reason that we label it as *conditional parametric bootstrap*. It has been used by Rao, Sinha and Dumitrescu (2014), and more recently by Chatrchi (2018) to estimate the MSE under a penalized spline mixed model.

In our context, for estimating $\text{MSE}\left(\hat{\bar{Y}}_i^{\text{LP}}\right)$, we proceed as follows. We generate $v_{1i}^* \sim N\left(0, \hat{\sigma}_{\text{glo},1v}^2\right)$ and $e_{1ij}^* \sim N\left(0, \hat{\sigma}_{\text{glo},1e}^2\right)$, and obtain the bootstrap responses

$$y_{1ij}^* = \tilde{\mathbf{x}}_{ij}^T\hat{\boldsymbol{\beta}}_{\text{glo},1} + \hat{m}_0\left(p_{j|i}\right) + v_{1i}^* + e_{1ij}^*,\ j \in U_i;\ i = 1, \ldots, M. \tag{4.3}$$

The $\hat{m}_0\left(p_{j|i}\right)$'s were estimated using the local model (3.6). The triplet $\left(\hat{\boldsymbol{\beta}}_{\text{glo},1}, \hat{\sigma}_{\text{glo},1v}^2, \hat{\sigma}_{\text{glo},1e}^2\right)$ was estimated using the global model (3.12) and the sample data $\left(y_{ij}, \tilde{\mathbf{x}}_{ij}, p_{j|i}\right)$, $j \in s_i$; $i = 1, \ldots, M$. The population bootstrap mean is $\bar{Y}_{1i}^* = N_i^{-1}\sum_{j=1}^{N_i} y_{1ij}^*$. Let $\hat{\boldsymbol{\beta}}_{\text{glo},1}^*$, $\hat{m}_0^*\left(p_{j|i}\right)$ and $\hat{v}_{\text{glo},1i}^*$ be bootstrap versions of estimators $\hat{\boldsymbol{\beta}}_{\text{glo},1}$, $\hat{m}_0\left(p_{j|i}\right)$ and $\hat{v}_{\text{glo},1i}$, that are based on bootstrap data $\left(y_{1ij}^*, \tilde{\mathbf{x}}_{ij}, p_{j|i}\right)$, $j \in s_i$; $i = 1, \ldots, M$ and the h_{opt} obtained with the original data set $\left(y_{ij}, \tilde{\mathbf{x}}_{ij}, p_{j|i}\right)$, $j \in s_i$; $i = 1, \ldots, M$. We did not re-compute the optimal h_{opt}^* associated with $\left(y_{1ij}^*, \tilde{\mathbf{x}}_{ij}, p_{j|i}\right)$, $j \in s_i$; $i = 1, \ldots, M$, as it would result in far too many computations in the Monte Carlo study. The bootstrap procedure is therefore conditional on $p_{j|i}$, $j \in U_i$; $i = 1, \ldots, M$ and h_{opt} obtained with the initial sample. Given that $\bar{s}_i$ is the set of non-sampled units in area i, the predicted bootstrap values $\hat{y}_{1ij}^*$ for $j \in \bar{s}_i$, are obtained as

$$\hat{y}_{1ij}^* = \tilde{\mathbf{x}}_{ij}^T\hat{\boldsymbol{\beta}}_{\text{glo},1}^* + \hat{m}_0^*\left(p_{j|i}\right) + \hat{v}_{\text{glo},1i}^*. \tag{4.4}$$

The resulting estimator of $\bar{Y}_{1i}^{*}$ is

$$\hat{\bar{Y}}_{1i}^{*} = \frac{1}{N_i}\left(\sum_{j \in s_i} y_{1ij}^{*} + \sum_{j \in \bar{s}_i} \hat{y}_{1ij}^{*}\right).$$

Repeating the above procedure B times, the conditional bootstrap estimator of MSE of the local polynomial estimator of $\bar{Y}_i$ is given by

$$\text{mse}_{\text{boot}}\left(\hat{\bar{Y}}_i^{\text{LP}}\right) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\bar{Y}}_{1i}^{*}(b) - \bar{Y}_{1i}^{*}(b)\right)^2, \tag{4.5}$$

where $\hat{\bar{Y}}_{1i}^{*}(b)$ and $\bar{Y}_{1i}^{*}(b)$ are the values of $\hat{\bar{Y}}_{1i}^{*}$ and $\bar{Y}_{1i}^{*}$ for the b^{th} bootstrap replicate.

The conditional bootstrap can also be used for estimating the mean squared error of an EBLUP estimator, $\hat{\bar{Y}}_i^{\text{VRH}}$, based on the augmented model (1.2) proposed by Verret et al. (2015). We included this procedure in the simulation given in Section 5, to get an idea of how the resulting MSE estimators compare to those obtained for $\hat{\bar{Y}}_i^{\text{LP}}$. The steps for obtaining the $\text{mse}\left(\hat{\bar{Y}}_i^{\text{VRH}}\right)$ are similar to those used for obtaining the mse of the local polynomial estimator $\hat{\bar{Y}}_i^{\text{LP}}$. In this case, bootstrap values for the responses y_{ij} are based on the augmented model (1.2) and the estimators $\left(\hat{\boldsymbol{\beta}}_0, \hat{\delta}_0\right)$ and $\left(\hat{\sigma}_{0v}^2, \hat{\sigma}_{0e}^2\right)$ obtained when the classical EBLUP theory is used with the sample data $\left(y_{ij}, \mathbf{x}_{ij}, g\left(p_{j|i}\right)\right)$, $j \in s_i$; $i = 1, \ldots, M$.

5 Simulation study

The set-up of the simulation study follows the one used in Verret et al. (2015). We considered a population with $M = 15$ small areas and $N_i = 15$ units within each small area. The relatively small number of small areas and units within areas were chosen so as to alleviate the computational burden. We used a single auxiliary variable x. The population x-values were generated from a gamma distribution with mean 10 and variance 50. The population y_{ij}-values were generated by the following model

$$y_{ij} = 4 + x_{ij} + v_i + e_{ij};\ i = 1, \ldots, 15;\ j = 1, \ldots, 15, \tag{5.1}$$

where $v_i \overset{\text{iid}}{\sim} N\left(0, \sigma_v^2\right)$ and $e_{ij} \overset{\text{iid}}{\sim} N\left(0, \sigma_e^2\right)$ with $\sigma_v^2 = 0.5$ and $\sigma_e^2 = 2$.

We considered a single sample size, $n_i = 3$, within a small area. We used Conditional Poisson Sampling (CPS) to select unequal probability samples within the small areas, with probabilities proportional to specified sizes c_{ij} (see Tillé, 2006, Chapter 5). We considered two different choices of the sizes c_{ij} in the simulation study. The first choice uses

$$c_{ij} = \exp\left[\frac{1}{3}\left\{-\frac{\left(v_i + e_{ij}\right)}{\sigma_e} + \frac{\delta_{ij}}{5}\right\}\right], \tag{5.2}$$

where $\delta_{ij} \overset{\text{iid}}{\sim} N(0, 1)$. The size measures (5.2) are equivalent to those used by Pfeffermann and Sverchkov (2007) in their simulation study and satisfy the relationship (2.5) on the weights $w_{j|i} = \pi_{j|i}^{-1}$.

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

$$c_{ij} = \left(1 + \exp\left\{-\tau\left(\frac{1}{\alpha}e_{ij} + \sqrt{1 - \frac{1}{\alpha^2}}\,e_{ij}^*\right)\right\}\right)^{-1}. \tag{5.3}$$

Non-invariant size measures are taken as

$$c_{ij} = \left(1 + \exp\left\{-\tau\left[\frac{1}{\alpha}\left(v_i + e_{ij}\right) + \sqrt{1 - \frac{1}{\alpha^2}}\left(v_i^* + e_{ij}^*\right)\right]\right\}\right)^{-1}, \tag{5.4}$$

where the random pair $\left(v_i^*, e_{ij}^*\right)$ is generated independently of $\left(v_i, e_{ij}\right)$ using the same distributions as v_i and e_{ij}. These size measures were used by Asparouhov (2006). The coefficient τ controls for the variation of the weights and the value α controls the level of informativeness of the sampling design. We chose $\tau = 0.5$ and $\alpha = 1, 2, 3$ and ∞ corresponding to several levels of informativeness generated by c_{ij} in (5.3) and (5.4). Increasing α decreases informativeness, with $\alpha = \infty$ corresponding to non-informative sampling. If some of the $\pi_{j|i}$'s exceeded one, they were set to one, and the probabilities were recomputed for the remaining units.

5.1 Performance of the local polynomial estimator of $\bar{Y}_i$

We compared the bias and mean squared error of the estimators $\hat{\bar{Y}}_i^{\text{EBLUP}}$, $\hat{\bar{Y}}_i^{\text{VRH}}$ and $\hat{\bar{Y}}_i^{\text{LP}}$. The EBLUP estimator $\hat{\bar{Y}}_i^{\text{EBLUP}}$ based on (1.1) assumes that the sample model coincides with the population model, thereby ignoring the informativeness of the sampling design. We studied two versions of $\hat{\bar{Y}}_i^{\text{VRH}}$ investigated by Verret et al. (2015) for various choices of $g(\cdot)$ that account for informativeness. They are EBLUP estimators based on the augmented sample model (1.2). They are denoted as $\hat{\bar{Y}}_i^{\text{VRH1}}$ when $g\left(p_{j|i}\right) = p_{j|i}$ and $\hat{\bar{Y}}_i^{\text{VRH2}}$ when $g\left(p_{j|i}\right) = \log\left(p_{j|i}\right)$. We report results only for these g functions, as they outperform others given in Verret et al. (2015). Finally, $\hat{\bar{Y}}_i^{\text{LP}}$ represents our new local polynomial estimator.

The bias and the mean squared error of the estimators were computed using $R = 1{,}000$ simulated samples selected under a design-model approach. For each run, $r = 1, \ldots, R$, we first generated the population y_{ij}-values under the population model (5.1) and computed $\bar{Y}_i^{(r)}$, the mean of the small area i in the r^{th} generated population. Samples of sizes $n_i = 3$ were then selected within the small areas using CPS with probabilities proportional to specified sizes $c_{ij}^{(r)}$ given by (5.2) for the Pfeffermann and Sverchkov (2007) (PS) size measures, and (5.3) and (5.4) corresponding to the invariant and non-invariant cases in the case of the Asparouhov (2006) (AP) size measures. From each simulated sample r $(r = 1, \ldots, R)$, the estimates $\hat{\bar{Y}}_i^{\text{EBLUP}(r)}$, $\hat{\bar{Y}}_i^{\text{VRH1}(r)}$, $\hat{\bar{Y}}_i^{\text{VRH2}(r)}$ and $\hat{\bar{Y}}_i^{\text{LP}(r)}$ were computed for each small area U_i. An optimal bandwidth $h_{\text{opt}}^{(r)}$ was found for $\hat{\bar{Y}}_i^{\text{LP}(r)}$ using the cross-validation criterion. A grid of the form (0.01, 0.02, 0.03, …, 0.15) covered the possible values for $h_{\text{opt}}^{(r)}$ in populations generated by (5.1).

For a given estimator of the small area mean $\bar{Y}_i$, we considered the following performance measures:

Average Absolute Bias

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

where

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

Average Root Mean Squared Error

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

Table 5.1 reports on the average absolute bias $\left(\overline{\text{AB}}\right)$ of estimators $\hat{\bar{Y}}_i^{\text{EBLUP}}$, $\hat{\bar{Y}}_i^{\text{VRH1}}$, $\hat{\bar{Y}}_i^{\text{VRH2}}$ and $\hat{\bar{Y}}_i^{\text{LP}}$ under the PS size measures (5.2) and AP size measures (5.3 and 5.4) for $\alpha = 1, 2, 3$ and ∞.

Table 5.1
Average absolute bias $\left(\overline{\text{AB}}\right)$ for the PS and AP size measures

Estimator Generation of $p_{j\mid i}$			$\hat{\bar{Y}}_i^{\text{EBLUP}}$ without $g\left(p_{j\mid i}\right)$	$\hat{\bar{Y}}_i^{\text{VRH1}}$ $g\left(p_{j\mid i}\right) = p_{j\mid i}$	$\hat{\bar{Y}}_i^{\text{VRH2}}$ $g\left(p_{j\mid i}\right) = \log\left(p_{j\mid i}\right)$	$\hat{\bar{Y}}_i^{\text{LP}}$ $m_0\left(p_{j\mid i}\right)$
PS			0.309	0.020	0.004	0.011
AP	$\alpha = 1$	I	0.431	0.002	0.036	0.004
		NI	0.425	0.010	0.035	0.005
	$\alpha = 2$	I	0.206	0.017	0.022	0.024
		NI	0.219	0.019	0.016	0.016
	$\alpha = 3$	I	0.139	0.005	0.012	0.033
		NI	0.137	0.008	0.013	0.019
	$\alpha = \infty$	I	0.008	0.008	0.008	0.026
		NI	0.006	0.006	0.006	0.021

As observed in Verret et al. (2015), the $\overline{\text{AB}}$ of the EBLUP estimator $\hat{\bar{Y}}_i^{\text{EBLUP}}$ with just the auxiliary variable x, is quite a bit larger than those based on the augmented models ($p_{j\mid i}$ and $\log(p_{j\mid i})$), and the local polynomial method. This holds regardless of how the size measures have been generated (PS or AP). The $\overline{\text{AB}}$ of $\hat{\bar{Y}}_i^{\text{EBLUP}}$ attains its highest value (0.431) when the design is very informative $(\alpha = 1)$, and decreases as α increases. This observation also holds for the estimators based on the augmented models. The inclusion of $p_{j\mid i}$ or $\log\left(p_{j\mid i}\right)$, as an augmenting variable, in the model results in small $\overline{\text{AB}}$'s, with the highest being 0.036. Comparing the $\overline{\text{AB}}$'s of the local polynomial estimator $\hat{\bar{Y}}_i^{\text{LP}}$ to those associated with the VRH augmented models, we observe that they are comparable for $\alpha = 1$ and $\alpha = 2$, and slightly larger for $\alpha \geq 3$.

Table 5.2 reports the simulation results on the average root mean squared error $\left(\overline{\text{RMSE}}\right)$ of the estimators for both the PS size measures (5.2) and the AP size measures (5.3 and 5.4) for $\alpha = 1, 2, 3$ and ∞. The EBLUP, $\hat{\bar{Y}}_i^{\text{EBLUP}}$, based on model (1.1) without the augmenting variable $g\left(p_{j\mid i}\right)$, has the largest $\overline{\text{RMSE}}$'s (0.740 for I and 0.752 for NI) for the AP size measures corresponding to $\alpha = 1$, and 0.685 for the PS size measure. The $\overline{\text{RMSE}}$ decreases as α increases: 0.608 for I and 0.610 for NI in the case of non-informative sampling $(\alpha = \infty)$. The $\overline{\text{RMSE}}$'s for $\hat{\bar{Y}}_i^{\text{VRH1}}$, $\hat{\bar{Y}}_i^{\text{VRH2}}$ and $\hat{\bar{Y}}_i^{\text{LP}}$ are significantly smaller than those associated with $\hat{\bar{Y}}_i^{\text{EBLUP}}$ when sampling is very informative $(\alpha = 1)$ and for the PS size measure. There are small differences in terms of $\overline{\text{RMSE}}$ between our non-parametric approach and the parametric approach in Verret et al. (2015).

Table 5.2
Average root mean squared error $\left(\overline{\textbf{RMSE}}\right)$ for the PS and AP size measures

Generation of $p_{j\mid i}$ \ Estimator			$\hat{\bar{Y}}_i^{\text{EBLUP}}$ without $g\left(p_{j\mid i}\right)$	$\hat{\bar{Y}}_i^{\text{VRH1}}$ $g\left(p_{j\mid i}\right) = p_{j\mid i}$	$\hat{\bar{Y}}_i^{\text{VRH2}}$ $g\left(p_{j\mid i}\right) = \log\left(p_{j\mid i}\right)$	$\hat{\bar{Y}}_i^{\text{LP}}$ $m_0\left(p_{j\mid i}\right)$
PS			0.685	0.229	0.200	0.200
AP	$\alpha = 1$	I	0.740	0.089	0.170	0.087
		NI	0.752	0.158	0.200	0.149
	$\alpha = 2$	I	0.644	0.562	0.568	0.557
		NI	0.650	0.557	0.555	0.555
	$\alpha = 3$	I	0.617	0.588	0.591	0.612
		NI	0.619	0.587	0.589	0.607
	$\alpha = \infty$	I	0.608	0.619	0.621	0.626
		NI	0.610	0.622	0.625	0.629

When the sampling is less informative $(\alpha = 3)$, the local linear estimator $\hat{\bar{Y}}_i^{\text{LP}}$ is better than $\hat{\bar{Y}}_i^{\text{EBLUP}}$, but its $\overline{\text{RMSE}}$ is slightly larger than those associated with the parametric estimators $\hat{\bar{Y}}_i^{\text{VRH1}}$ and $\hat{\bar{Y}}_i^{\text{VRH2}}$. In this case, we observe that the estimated function $m_0\left(p_{j\mid i}\right)$ is close to a flat line, and this implies that the local linear approximation is not as appropriate. This explains why $\hat{\bar{Y}}_i^{\text{LP}}$ is slightly worse than $\hat{\bar{Y}}_i^{\text{VRH1}}$ and $\hat{\bar{Y}}_i^{\text{VRH2}}$ when the level of informativeness of the sampling is low. A local polynomial estimator performs well when the function $m_0(\cdot)$ is meaningfully non-constant.

When the sample is non-informative $(\alpha = \infty)$, $\hat{\bar{Y}}_i^{\text{EBLUP}}$ is better than $\hat{\bar{Y}}_i^{\text{VRH1}}$, $\hat{\bar{Y}}_i^{\text{VRH2}}$ and $\hat{\bar{Y}}_i^{\text{LP}}$ in both invariant and non-invariant case. This conclusion is somewhat different from that of Verret et al. (2015) where for $\alpha = \infty$ their estimators $\hat{\bar{Y}}_i^{\text{EBLUP}}$, $\hat{\bar{Y}}_i^{\text{VRH1}}$ and $\hat{\bar{Y}}_i^{\text{VRH2}}$ have equal $\overline{\text{AB}}$ and $\overline{\text{RMSE}}$ values. Verret et al. (2015) used both larger populations and samples, and this may explain why their augmented models produced estimators as good as the population model under non-informative sampling designs. Under our simulation set-up, we found that the $\overline{\text{AB}}$ and $\overline{\text{RMSE}}$ of the EBLUP are small for α values larger than 6: this corresponds to a sample design that is almost non-informative. In this case, we recommend using EBLUP.

5.2 Performance of the MSE estimators

We now turn to the performance of the bootstrap procedures for estimating the MSEs of the EBLUP, VRH and local polynomial estimators. Let $\hat{\bar{Y}}_i$ be an estimator of $\bar{Y}_i$ and $\text{mse}_{\text{boot}}\left(\hat{\bar{Y}}_i\right)$ be the bootstrap estimator of $\text{MSE}\left(\hat{\bar{Y}}_i\right)$. From $R = 1{,}000$ simulated populations and samples, we first computed measures of MSE values as

$$\text{MSE}\left(\hat{\bar{Y}}_i\right) = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\bar{Y}}_i^{(r)} - \bar{Y}_i^{(r)}\right)^2,$$

where $\bar{Y}_i^{(r)}$ is the true mean, and $\hat{\bar{Y}}_i^{(r)}$ is the value of the estimator for the r^{th} population. Let $\text{mse}_{\text{boot}}\left(\hat{\bar{Y}}_i\right)$ be the bootstrap estimator of $\text{MSE}\left(\hat{\bar{Y}}_i\right)$. It is denoted as $\text{mse}_{\text{boot}}\left(\hat{\bar{Y}}_i^{\text{EBLUP}}\right)$ for the EBLUP estimator $\hat{\bar{Y}}_i^{\text{EBLUP}}$, and corresponds to the parametric (unconditional) bootstrap method given by equation (4.2). For our local polynomial estimator $\hat{\bar{Y}}_i^{\text{LP}}$ and the Verret et al. (2015) estimators, $\hat{\bar{Y}}_i^{\text{VRH1}}$ and $\hat{\bar{Y}}_i^{\text{VRH2}}$, the mse values, denoted as $\text{mse}_{\text{boot}}\left(\hat{\bar{Y}}_i^{\text{LP}}\right)$ and $\text{mse}_{\text{boot}}\left(\hat{\bar{Y}}_i^{\text{VRH}j}\right)$, for $j = 1$ and $j = 2$ respectively, are computed using the conditional parametric bootstrap method of Section 4. For each selected sample in the r^{th} simulated population $(r = 1, \ldots, R)$, we used $B = 400$ bootstraps to compute the r^{th} value of $\text{mse}_{\text{boot}}\left(\hat{\bar{Y}}_i\right)$, that we denote as $\text{mse}_{\text{boot}}^{(r)}\left(\hat{\bar{Y}}_i\right)$. We considered two measures to evaluate the performance of $\text{mse}_{\text{boot}}\left(\hat{\bar{Y}}_i\right)$: average absolute relative bias and average confidence interval. These measures are defined as follows:

Average Absolute Relative Bias:

$$\overline{\text{ARB}} = \frac{1}{M}\sum_{i=1}^{M}\left|\frac{E\left(\text{mse}_{\text{boot}}\left(\hat{\bar{Y}}_i\right)\right)}{\text{MSE}\left(\hat{\bar{Y}}_i\right)} - 1\right|,$$

where

$$E\left(\text{mse}_{\text{boot}}\left(\hat{\bar{Y}}_i\right)\right) = \frac{1}{R}\sum_{r=1}^{R}\text{mse}_{\text{boot}}^{(r)}\left(\hat{\bar{Y}}\right).$$

Average Confidence Level:

$$\overline{\text{CL}} = \frac{1}{M}\sum_{i=1}^{M}\text{CL}_i,$$

where $\text{CL}_i = R^{-1}\sum_{r=1}^{R}\text{I}\left(\bar{Y}_i^{(r)} \in \text{IC}^{(r)}\right)$ and $\text{IC}^{(r)} = \left[\hat{\bar{Y}}_i^{(r)} \pm 1.96\sqrt{\text{mse}_{\text{boot}}^{(r)}\left(\hat{\bar{Y}}_i\right)}\,\right]$.

Table 5.3 reports simulation results on the average relative bias $\left(\overline{\text{ARB}}\right)$ of the MSE estimators for both the PS size measures (5.2) and Asparouhov size measures (5.3 and 5.4) for $\alpha = 1, 2, 3$ and ∞.

Table 5.3
Average relative bias (%) of mse $\left(\overline{\text{ARB}}\right)$ for the PS and AP size measures

Estimator Generation of $p_{j\mid i}$			$\hat{\bar{Y}}_i^{\text{EBLUP}}$ without $g\left(p_{j\mid i}\right)$	$\hat{\bar{Y}}_i^{\text{VRH1}}$ $g\left(p_{j\mid i}\right) = p_{j\mid i}$	$\hat{\bar{Y}}_i^{\text{VRH2}}$ $g\left(p_{j\mid i}\right) = \log\left(p_{j\mid i}\right)$	$\hat{\bar{Y}}_i^{\text{LP}}$ $m_0\left(p_{j\mid i}\right)$
PS			25.4	3.9	3.4	7.7
AP	$\alpha = 1$	I	39.9	9.7	14.4	7.5
		NI	46.6	4.1	8.7	10.0
	$\alpha = 2$	I	16.0	2.9	3.8	5.9
		NI	21.4	3.8	3.5	5.8
	$\alpha = 3$	I	13.4	6.1	6.4	5.8
		NI	15.4	7.3	7.4	8.8
	$\alpha = \infty$	I	4.6	4.2	4.5	6.2
		NI	6.1	6.4	6.3	6.9

The $\overline{\text{ARB}}$ of $\hat{\bar{Y}}_i^{\text{EBLUP}}$, based on the model without the augmenting variable $g\left(p_{j|i}\right)$, is very large when the sampling is very informative $(\alpha = 1)$: 39.9% for I and 46.6% for NI. The $\overline{\text{ARB}}$ gradually decreases to around 5% under non-informative sampling $(\alpha = \infty)$. The $\overline{\text{ARB}}$'s of both the parametric and non-parametric estimators are smaller in general than 10%, with the exception of 14.4% for the $\hat{\bar{Y}}_i^{\text{VRH2}}$ estimator that uses $\log\left(p_{j|i}\right)$ as an augmenting variable.

Table 5.4 reports simulation results on the average confidence level $\left(\overline{\text{CL}}\right)$ associated with the MSE estimators for both the PS size measures (5.2) and the AP size measures (5.3 and 5.4) for $\alpha = 1, 2, 3$ and ∞ and nominal level of 0.95.

Table 5.4
Average confidence level of mse $\left(\overline{\text{CL}}\right)$ for the PS and AP size measures

Estimator Generation of $p_{j\|i}$			$\hat{\bar{Y}}_i^{\text{EBLUP}}$ without $g\left(p_{j\|i}\right)$	$\hat{\bar{Y}}_i^{\text{VRH1}}$ $g\left(p_{j\|i}\right) = p_{j\|i}$	$\hat{\bar{Y}}_i^{\text{VRH2}}$ $g\left(p_{j\|i}\right) = \log\left(p_{j\|i}\right)$	$\hat{\bar{Y}}_i^{\text{LP}}$ $m_0\left(p_{j\|i}\right)$
PS			0.898	0.937	0.941	0.936
AP	$\alpha = 1$	I	0.856	0.918	0.908	0.928
		NI	0.834	0.930	0.920	0.934
	$\alpha = 2$	I	0.916	0.937	0.936	0.932
		NI	0.907	0.936	0.933	0.936
	$\alpha = 3$	I	0.922	0.927	0.926	0.934
		NI	0.918	0.930	0.933	0.926
	$\alpha = \infty$	I	0.937	0.935	0.935	0.938
		NI	0.934	0.934	0.933	0.931

The EBLUP estimator $\hat{\bar{Y}}_i^{\text{EBLUP}}$ has the worst coverage when the sample design is very informative. The coverage improves as the design becomes less informative. The coverage of the other estimators is between 93% and 95%, with the exception of $\hat{\bar{Y}}_i^{\text{VRH2}}$ (the one that includes $\log(p_{j|i})$) with coverage slightly lower.

5.3 Inclusion of an augmenting variable

The local polynomial approach results in an automatic way of obtaining a reasonable augmented model that is a function of the selection probabilities $p_{j|i}$. However, given that one does not know whether the design is informative or not, should we always include an augmenting variable in the model? If the sample design is not informative it is reasonable to use model (1.1). Note that in this case, including the augmenting variables, $p_{j|i}$ or $\log\left(p_{j|i}\right)$, has a very small impact either on the absolute relative bias of the estimator and absolute relative bias of the estimated MSE. A similar conclusion was obtained in Verret et al. (2015) who used a larger population and sample size.

The same question arises with respect to the use of the local polynomial procedure. In this case, the conclusions are not quite as clear. If the design is very informative, the local polynomial approach gains in terms of absolute bias and mean squared error when $\alpha = 1$ or $\alpha = 2$. When the sampling design is less informative $(\alpha = 3)$ the parametric approach in Verret et al. (2015) is the better choice, but by a very small margin.

In a practical situation, the value of α is not known and the decision to use the augmenting variable in a parametric or nonparametric model should be taken. To this end, we follow the suggested procedure in Verret et al. (2015) to provide some guidelines on how to decide on this choice for an arbitrary data set. Define $u_{ij} = v_i + e_{ij}$, and fit the following model $y_{ij} = \beta_0 + \beta_1 x_{ij} + u_{ij}$ to the sample data by ordinary least squares (OLS). The residuals are $\tilde{u}_{ij} = y_{ij} - \tilde{\beta}_0 - \tilde{\beta}_1 x_{ij}$, where $\tilde{\beta}_0$ and $\tilde{\beta}_1$ are the OLS estimators of β_0 and β_1 respectively. Figure 5.1 displays residual plots of $\left(\hat{m}_0\left(p_{j|i}\right), \tilde{u}_{ij}\right)$, $j = 1, \ldots, N_i$; $i = 1, \ldots, M$ for the AP measures $\alpha = 1, 2, 3$ and ∞ in the invariant case. For $\alpha = 1$, the relationship between $\tilde{u}_{ij}$ and $\hat{m}_0\left(p_{j|i}\right)$ is clearly linear, suggesting that the design is informative. As α increases, the design is less informative. Note that $m_0\left(p_{j|i}\right)$ is constant when $\alpha = \infty$. Similar observations hold for the non-invariant case. For the PS size measures the graph resembled the one given in Figure 5.1 when $\alpha = 1$.

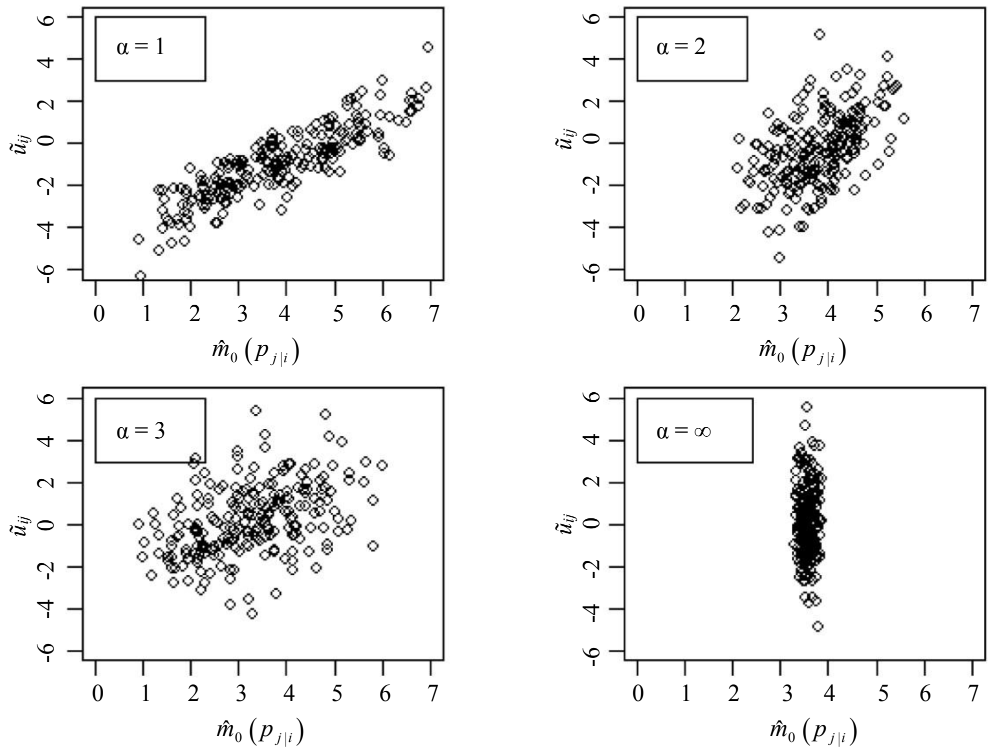

Figure 5.1 Residual plots for the population: AP invariant size measures.

Table 5.5 provides the estimated correlation coefficients, $\hat{\rho} = \operatorname{cor}\left(\tilde{u}_{ij}, \hat{m}_0\left(p_{j|i}\right)\right)$, for PS and AP size measures for $\alpha = 1, 2, 3$ and ∞.

Table 5.5
Estimated correlation coefficient $\hat{\rho} = \operatorname{cor}\left(\tilde{u}_{ij}, \hat{m}_0\left(p_{j|i}\right)\right)$ for the PS and AP size measures

Estimated correlation coefficient	AP								PS
	$\alpha = 1$		$\alpha = 2$		$\alpha = 3$		$\alpha = \infty$		
	I	NI	I	NI	I	NI	I	NI	
$\hat{\rho}$	0.870	0.850	0.450	0.510	0.240	0.210	0.007	0.001	0.800

In terms of $\overline{\text{RMSE}}$, we noticed in Section 5.1 that $\hat{\bar{Y}}_i^{\text{EBLUP}}$ is better than the estimators based on augmented models for $\alpha \geq 6$. Results not presented in Table 5.5 show that for $\alpha \geq 6$, the absolute value of the correlation coefficient is less than 0.1. On the basis of this limited simulation, a user could decide on the choice of the estimator to use for a real data set as follows: i. If $|\hat{\rho}|$ is larger than 0.5, use $\hat{\bar{Y}}_i^{\text{LP}}$; ii. If $|\hat{\rho}|$ is less than 0.1, use $\hat{\bar{Y}}_i^{\text{EBLUP}}$; iii. otherwise use $\hat{\bar{Y}}_i^{\text{VRH1}}$ or $\hat{\bar{Y}}_i^{\text{VRH2}}$.

6 Concluding remarks

In this paper, we studied the estimation of a small area mean under informative sampling by using an augmented model approach where the augmenting variable is a smooth function $m_0\left(p_{j|i}\right)$ of the selection probability $p_{j|i}$. Our augmented model is semi-parametric. It differs from Verret et al. (2015), in that nothing was assumed about the augmenting function $m_0(\cdot)$.

We proposed a three-step procedure to estimate the augmented semi-parametric model. Firstly, local polynomial fits were estimated for each unit of the population (sampled and non-sampled). Secondly, given these local fits a new dependent variable was defined to obtain global estimators of the regression parameters and the small area effects. The resulting estimators were used to compute the predicted values of the dependent variable, y, for all non-sampled units. Finally, using the observed sample values of y, and the predicted values of y, we computed the local polynomial estimator $\hat{\bar{Y}}_i^{\text{LP}}$ for the small area mean $\bar{Y}_i$.

We adopted the conditional parametric bootstrap method to estimate the mean squared error of the newly proposed estimator. The conditional bootstrap is a modified version of the parametric bootstrap estimator method of Hall and Maiti (2006).

We carried out a simulation study to compare the bias and mean squared error performance of the usual EBLUP, $\hat{\bar{Y}}_i^{\text{EBLUP}}$, the augmented EBLUP of Verret et al. (2015), $\hat{\bar{Y}}_i^{\text{VRH}}$, and the proposed local polynomial estimator, $\hat{\bar{Y}}_i^{\text{LP}}$. As expected, $\hat{\bar{Y}}_i^{\text{EBLUP}}$ exhibited large bias under informative sampling. The new estimator $\hat{\bar{Y}}_i^{\text{LP}}$ had equal or smaller MSE than $\hat{\bar{Y}}_i^{\text{VRH}}$ when the sample design was highly informative. If the sample design is less informative, it is better to use one of the two estimators in Verret et al. (2015): that is, augment the basic model with either $p_{j|i}$ or $\log\left(p_{j|i}\right)$. Note that in doing so, the gains are very small. If the sampling design is very slightly or not at all informative, then estimator $\hat{\bar{Y}}_i^{\text{EBLUP}}$ based on the population model should be used.

We also evaluated the performance of the mean squared error bootstrap estimation for the estimators $\hat{\bar{Y}}_i^{\text{EBLUP}}$, $\hat{\bar{Y}}_i^{\text{VRH}}$ and $\hat{\bar{Y}}_i^{\text{LP}}$, in terms of average absolute relative bias $\left(\overline{\text{ARB}}\right)$ and average confidence level $\left(\overline{\text{CL}}\right)$. The conditional bootstrap provides a good way to estimate the mean squared errors.

The advantage of the local polynomial approach is that it provides an automatic way of augmenting the model when the design is informative. Its biggest disadvantage is its computational burden both in terms of parameter estimation and associated reliability. The procedure outlined in Section 5.3 suggests a way to determine whether it is worth using it or not. An alternative approach is to augment the unit level model with a P-spline term of selection probabilities to account for the informativeness of the sampling design. This approach has been recently studied by Chatrchi (2018).

Acknowledgements

We would like to thank J.N.K. Rao for suggesting the conditional bootstrap for estimating the mean squared error of the polynomial estimator, and commenting on the article. We would also like to thank the Associate Editor and a referee for their constructive comments that have improved the paper.

References

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

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

Breidt, F.J., and Opsomer, J.D. (2000). Local polynomial regression estimators in survey sampling. *Annals of Statistics*, 28, 4, 1026-1053.

Breidt, F.J., Opsomer, J.D., Johnson, A.A. and Ranalli, M.G. (2007). Semiparametric model-assisted estimation for natural resource surveys. *Survey Methodology*, 33, 1, 35-44. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2007001/article/9850-eng.pdf.

Chatrchi, G. (2018). *Small Area Estimation: Informative Sampling and Two-Fold Models*. Unpublished Ph.D. Thesis, Carleton University, Ottawa, Canada.

Fan, J., and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications*.London: Chapman and Hall.

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, 5, 443-462.

Hall, P., and Maiti, T. (2006). On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 68, 221-238.

Opsomer, J.D., Claeskens, G., Ranalli, M.G., Kauermann, G. and Breidt, F.J. (2008). Nonparametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1), 265-286.

Pfeffermann, D., and Sverchkov, M. (2007). Small-area estimation under informative probability sampling of areas and within the selected areas. *Journal of the American Statistical Association*, 102, 480, 1427-1439.

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

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

Rao, J.N.K., Sinha, S.K. and Dumitrescu, L. (2014). Robust small area estimation under semi-parametric mixed models. *The Canadian Journal of Statistics*, 42, 126-141.

Ruppert, D., and Matteson, D.E. (2015). *Statistics and Data Analysis for Financial Engineering with R Examples: 2nd Ed.*, New York: Springer.

Tillé, Y. (2006). *Sampling Algorithms:* New York: Springer.

Verret, F., Rao, J.N.K. and Hidiroglou, M.A. (2015). Model-based small area estimation under informative sampling. *Survey Methodology*, 41, 2, 333-347. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2015002/article/14248-eng.pdf.

Wu, H., and Zhang, J.T. (2002). Local polynomial mixed-effects models for longitudinal data. *Journal of the American Statistical Association*, 97, 459, 883-897.

Survey Methodology, June 2020
Vol. 46, No. 1, pp. 51-75
Statistics Canada, Catalogue No. 12-001-X

Small area estimation methods under cut-off sampling

María Guadarrama, Isabel Molina and Yves Tillé[1]

Abstract

Cut-off sampling is applied when there is a subset of units from the population from which getting the required information is too expensive or difficult and, therefore, those units are deliberately excluded from sample selection. If those excluded units are different from the sampled ones in the characteristics of interest, naïve estimators may be severely biased. Calibration estimators have been proposed to reduce the design-bias. However, when estimating in small domains, they can be inefficient even in the absence of cut-off sampling. Model-based small area estimation methods may prove useful for reducing the bias due to cut-off sampling if the assumed model holds for the whole population. At the same time, for small domains, these methods provide more efficient estimators than calibration methods. Since model-based properties are obtained assuming that the model holds but no model is exactly true, here we analyze the design properties of calibration and model-based procedures for estimation of small domain characteristics under cut-off sampling. Our results confirm that model-based estimators reduce the bias due to cut-off sampling and perform significantly better in terms of design mean squared error.

Key Words: Calibration estimators; Cut-off sampling; Empirical best linear unbiased predictor (EBLUP); Empirical best/Bayes predictor (EBP); Nested-error model; Unit level models.

1 Introduction

Haziza, Chauvet and Deville (2010) describe cut-off sampling as a technique in which a set of units is deliberately excluded from possible selection in the sample. For the Organisation for Economic Co-operation and Development (OECD), it is a sampling procedure in which a threshold is established such that all units above or below the threshold are excluded from selection in a sample. According to Särndal, Swensson and Wretman (1992, pages 531-533), this sampling technique is typically used when the distribution of the study variable is highly skewed and there is no reliable frame covering the small elements. Benedetti, Bee and Espa (2010) recognizes the advantage of cut-off sampling in terms of survey reduction cost. This procedure is often used in business surveys, where small firms are deliberately excluded from the sample due to difficulty of getting information from them. The cost of obtaining and keeping a reliable frame for the whole population does not compensate the subsequent gain in accuracy.

The monthly survey of manufacturing performed by Statistics Canada is an example of cut-off sampling (Benedetti et al., 2010). In Spain, the monthly survey of industrial production index (IPI) performed by the Spanish National Statistical Institute (in Spanish, INE) collects data from firms that produce a significant volume of products according to the annual industrial survey of products (in Spanish EIAP), see INE (2018). Related surveys, e.g., the index of industrial prices (IIP) and the index of business turnover (IBT) also use one form of cut-off sampling. Since the inclusion probabilities for the excluded units are zero, this procedure leads to biased design-based estimators, see e.g., Särndal et al. (1992) or Haziza et al. (2010) among others. To reduce the cut-off sampling bias, Haziza et al. (2010) propose to use

[1] María Guadarrama, Luxembourg Institute of Socio-Economic Research (LISER), 11, Porte des Sciences, Campus Belval L-4366 Esch-sur-Alzette, Luxembourg. E-mail: maria.guadarrama@liser.lu; Isabel Molina, Universidad Carlos III de Madrid, C/Madrid 126, 28903, Getafe, Madrid, Spain. E-mail: isabel.molina@uc3m.es; Yves Tillé, Institut de Statistique, Université de Neuchâtel, 51, Av. de Bellevaux, 2000 Neuchâtel, Switzerland. E-mail: yves.tille@unine.ch.

auxiliary information either at the design or at the estimation stage; concretely, they propose to use balanced sampling and/or calibration.

In this work, we restrict ourselves to the estimation stage and study how cut-off sampling affects the estimation of domain (or area) parameters. We analyze some of the calibration methods proposed by Haziza et al. (2010) to reduce this problem. For domains with small sample size (small domains or areas), even in absence of cut-off sampling, calibration estimators might be inefficient. To improve efficiency, we consider small area estimation methods. For estimation of linear parameters, we consider the empirical best linear unbiased predictor (EBLUP) and, for general non-linear parameters, we consider the empirical best/Bayes predictor (EBP). We apply the methods studied in this work to the estimation of the total sales of certain tobacco product in Spanish provinces.

In the absence of cut-off sampling, the considered model-based estimators are approximately optimal when the model holds for all the population units. However, since no model holds exactly, we wish to study whether model-based estimators still perform better than basic design-based estimators (which do not depend on models) and calibration estimators under the sampling replication mechanism; i.e., without model assumptions and when cut-off sampling is present.

The article is organized as follows. Section 2 describes the theoretical set-up. The following four sections describe the considered estimation methods, namely the basic direct estimators (Section 3), different approaches to calibration (Section 4), the EBLUP for estimation of linear parameters (Section 5) and the EBP for estimation of more general parameters in small domains (Section 6). Section 7 describes a bootstrap procedure for estimating the mean squared error of the proposed small area estimators. Section 8 compares, through simulation experiments, the performance of several small area estimators under cut-off sampling. Section 9 describes the application and, finally, Section 10 draws some conclusions.

2 Cut-off sampling in small domains

We consider a population U partitioned into m subsets $U_i, i = 1, \ldots, m,$ called hereafter domains or areas, of sizes $N_i, i = 1, \ldots, m,$ with $N = \sum_{i=1}^{m} N_i.$ We restrict ourselves to the case in which the domains act as sampling strata. Then, independent samples are drawn from the different domains, where the sample s_i of size n_i from domain i is supposed to be drawn by cut-off sampling, $i = 1, \ldots, m.$ This is done by excluding a subset of units $U_{iE} \subseteq U_i$ from the selection. In other words, the domain U_i is partitioned into two subsets, U_{iI} and U_{iE}, of known sizes N_{iI} and N_{iE} respectively, with $N_i = N_{iI} + N_{iE}$. The set U_{iI} contains the units that can be potentially selected for the sample, called here the set of included units, whereas U_{iE} contains the excluded units.

Let y_{ij} be the value of the target variable y for the j^{th} unit within the i^{th} domain. We focus on estimation of domain totals $Y_i = \sum_{i=1}^{N_i} y_{ij}$ or means $\bar{Y}_i = Y_i / N_i$, $i = 1, \ldots, m.$ Under cut-off sampling within each domain, the sample s_i is supposed to be drawn from the subset of included individuals, U_{iI}, from domain i. Then, the inclusion probabilities for the included individuals $(j \in U_{iI})$ are $\pi_{j|i} = \Pr(j \in s_i) > 0$ and $w_{j|i} = \pi_{j|i}^{-1}$ are the corresponding sampling weights. For the excluded units

$(j \in U_{iE})$, the inclusion probabilities are zero and, therefore, the corresponding sampling weights are not defined. As a consequence, for domains i with $U_{iE} \neq \varnothing$, basic design-based estimators of Y_i or $\bar{Y}_i$ are biased and a design-unbiased estimator does not exist.

3 Basic direct estimators

We first consider basic direct estimators, obtained using only the n_i observations of the variable of interest from the target area. In the absence of cut-off sampling, these estimators are design-consistent as the domain sample size n_i increases. Moreover, they are nonparametric in the sense that do not require any model assumption. However, they may have unacceptable sampling errors in small domains. In addition, as we shall see below, under cut-off sampling, their design-bias might be substantial.

The usual expansion estimator (Horvitz and Thompson, 1952) of Y_i obtained ignoring that the sample s_i is drawn only from U_{iI} is given by $\hat{Y}_i = \sum_{j \in s_i} w_{ij} y_{ij}$. Under cut-off sampling, $\hat{Y}_i$ actually estimates the total in the included strata, $Y_{iI} = \sum_{i \in U_{iI}} y_{ij}$, rather than the overall total $Y_i = Y_{iI} + Y_{iE}$, where $Y_{iE} = \sum_{i \in U_{iE}} y_{ij}$. Indeed, $E_\pi(\hat{Y}_i) = Y_{iI}$, where E_π denotes expectation under repeated sampling, since the sampling weights $w_{j|i} = \pi_{j|i}^{-1}$ in $\hat{Y}_i$ expand to U_{iI} instead of U_i. No one would use this estimator since its bias, $B_\pi(\hat{Y}_i) = E_\pi(\hat{Y}_i) - Y_i = -Y_{iE}$, given in relative terms by the proportion of the total represented by the excluded population, $\mathrm{RB}_\pi(\hat{Y}_i) = -Y_{iE} / Y_i$, can be substantial.

When auxiliary information is not available, it makes more sense to use the Hájek estimator (Hájek, 1971) for the mean $\bar{Y}_i$, given by $\hat{\bar{Y}}_i^{\mathrm{HA}} = \hat{Y}_i / \hat{N}_i$, where $\hat{N}_i = \sum_{j \in s_i} w_{ij}$. The corresponding estimator for the total is $\hat{Y}_i^{\mathrm{HA}} = N_i \hat{\bar{Y}}_i^{\mathrm{HA}}$, considering that the means in the included and excluded strata are equal. Indeed, ignoring the ratio bias (of lower order) and noting that $E_\pi(\hat{N}_i) = N_{iI}$, the asymptotic (as $n_i \to \infty$) design-bias of $\hat{Y}_i^{\mathrm{HA}}$ is given in absolute and relative terms by

$$B_\pi(\hat{Y}_i^{\mathrm{HA}}) \cong N_{iE}(\bar{Y}_{iI} - \bar{Y}_{iE}), \qquad \mathrm{RB}_\pi(\hat{Y}_i^{\mathrm{HA}}) \cong \frac{N_{iE}}{N_i} \frac{\bar{Y}_{iI} - \bar{Y}_{iE}}{\bar{Y}_i}, \tag{3.1}$$

where $\bar{Y}_{iI} = Y_{iI} / N_{iI}$ and $\bar{Y}_{iE} = Y_{iE} / N_{iE}$ are the true means of the sets of included and excluded units from area i respectively (Haziza et al., 2010). For the mean, the bias of $\hat{\bar{Y}}_i^{\mathrm{HA}}$ is obtained dividing by N_i in (3.1). For a domain i with $U_{iE} \neq \varnothing$, the above bias vanishes only when $\bar{Y}_{iI} = \bar{Y}_{iE}$, which is unlikely in the real cases where cut-off sampling is applied, see e.g., Haziza et al. (2010) or Section 9. In the next section, we briefly describe calibration techniques as a mean of reducing the cut-off sampling bias.

Remark 3.1. The Hájek estimator of $\bar{Y}_i$ is a special case of the customary ratio estimator. In many monthly business surveys, parameters of interest are actually the changes over time of certain totals, such as $\theta_{it} = Y_i(t) / Y_i(t-1)$, where $Y_i(t)$ is the total of the target variable at time t within domain i. The ratio estimates of change are actually reported instead of the actual totals because it is often believed that such ratios are not affected by cut-off sampling bias. Let $\hat{\theta}_{it} = \hat{Y}_i(t) / \hat{Y}_i(t-1)$ be the basic direct estimator of θ_{it}. As we have seen above, the bias of the ratio estimator due to cut-off sampling tends to be much smaller than that of the absolute totals $\hat{Y}_i(t)$ and $\hat{Y}_i(t-1)$. However, as we have also seen, the cut-off sampling bias of ratio estimators vanishes only under strong assumptions. Indeed, ignoring the ratio bias, which is negligible for large n_i, the bias of $\hat{\theta}_{it}$ is given by

$$B_\pi\left(\hat{\theta}_{it}\right) \cong \frac{Y_{iI}(t)}{Y_{iI}(t-1)} - \frac{Y_i(t)}{Y_i(t-1)},$$

where $Y_{iI}(t)$ denotes the corresponding total for the included units only. This bias is zero only if the ratios for the population $Y_i(t)/Y_i(t-1)$ are the same as those for the included units $Y_{iI}(t)/Y_{iI}(t-1)$.

4 Calibration estimators

Calibration is traditionally applied when the true totals of certain auxiliary variables, which are potentially correlated with the study variable, are known. The idea of calibration is to adjust the design weights $w_{j|i}$, so that the corresponding expansion estimators of the available true totals have zero error. If the adjusted weights provide estimators of the available totals of the auxiliary variables that are absent of error, then one expects that they will also decrease the error in the estimation of the total of the study variable, provided that it is linearly related with the auxiliary variables. Even if there is an underlying linear model, in the absence of cut-off sampling, calibration estimators are design-consistent as the area sample size n_i increases even if the model does not hold. In this sense, they are model-assisted and their properties are typically evaluated under the design-based setup. However, if n_i is small, the estimates may suffer from small sample bias.

As we shall see below, calibration estimators reduce the bias due to cut-off sampling if the underlying linear model holds for the whole population (included and excluded units). However, for small domains, they might have unacceptably large sampling errors, apart from non-negligible small sample bias.

Let us denote by $\mathbf{x}_{ij}$ the vector of auxiliary variables for unit j within domain i. Depending on whether the domain totals or only the population totals of these auxiliary variables are available, we can apply different calibration approaches. First, consider the case whereby the vector of domain totals $\mathbf{X}_i = \sum_{j=1}^{N_i} \mathbf{x}_{ij}$ is available. Note that $\mathbf{X}_i$ is the total in the whole domain $U_i = U_{iI} \cup U_{iE}$. Then, one approach to calibration is to determine calibration weights $h_{j|i}$, $j \in s_i$, that minimize

$$\sum_{j \in s_i} \left(h_{j|i} - w_{j|i}\right)^2 \Big/ w_{j|i} \qquad (4.1)$$

$$\text{s.t.} \sum_{j \in s_i} h_{j|i} \mathbf{x}_{ij} = \mathbf{X}_i.$$

The resulting calibration weights $h_{j|i}$ are given by

$$h_{j|i} = w_{j|i} \left\{ 1 + \left(\mathbf{X}_i - \hat{\mathbf{X}}_i\right)' \left(\sum_{j \in s_i} w_{j|i} \mathbf{x}_{ij} \mathbf{x}'_{ij} \right)^{-1} \mathbf{x}_{ij} \right\}, \ j \in s_i, \qquad (4.2)$$

provided that $\sum_{j \in s_i} w_{j|i} \mathbf{x}_{ij} \mathbf{x}'_{ij}$ is non-singular. The calibration estimator of the domain total Y_i is then given by

$$\hat{Y}_i^{\mathrm{LCAL}} = \sum_{j \in s_i} h_{j|i} y_{ij} = \hat{Y}_i + \left(\mathbf{X}_i - \hat{\mathbf{X}}_i\right)' \hat{\mathbf{B}}_i, \tag{4.3}$$

which is the well-known generalized regression (GREG) estimator of Y_i, where

$$\hat{\mathbf{B}}_i = \left(\sum_{j \in s_i} w_{j|i} \mathbf{x}_{ij} \mathbf{x}'_{ij}\right)^{-1} \sum_{j \in s_i} w_{j|i} \mathbf{x}_{ij} y_{ij}.$$

The Hájek estimator $\hat{Y}_i^{\mathrm{HA}}$ is a special case of (4.3), with $\mathbf{x}_{ij} = 1,\ j = 1, \ldots, N_i$. In the absence of cut-off sampling, the above GREG estimator is design-consistent as the domain sample size n_i increases, although it may suffer from small sample bias. It reduces the variance if the calibration variables are linearly correlated with the outcome and the correlation is strong. Under cut-off sampling, the second term on the right-hand side of (4.3) corrects for the bias of the basic expansion estimator $\hat{Y}_i$ as estimator of Y_i with the help of the known domain totals in $\mathbf{X}_i$. However, for small domain sample size n_i, this reduction in cut-off sampling bias might be transferred to an increase in variance.

In the above procedure, we have a different calibration problem for each domain. In the case that only the overall population total $\mathbf{X} = \sum_{i=1}^m \sum_{j=1}^{N_i} \mathbf{x}_{ij}$ is available, we may seek calibration weights for all the domains at once, $g_{j|i},\ j \in s_i,\ i = 1, \ldots, m,$ by solving only one calibration problem:

$$\min_{\{g_{j|i}:\ j \in s_i,\ i = 1, \ldots, m\}} \sum_{i=1}^m \sum_{j \in s_i} \left(g_{j|i} - w_{j|i}\right)^2 \Big/ w_{j|i} \tag{4.4}$$

$$\text{s.t.} \sum_{i=1}^m \sum_{j \in s_i} g_{j|i} \mathbf{x}_{ij} = \mathbf{X}.$$

In this case, the calibration weights $g_{j|i}$ are given by

$$g_{j|i} = w_{j|i} \left\{1 + \left(\mathbf{X} - \hat{\mathbf{X}}\right)' \left(\sum_{i=1}^m \sum_{j \in s_i} w_{j|i} \mathbf{x}_{ij} \mathbf{x}'_{ij}\right)^{-1} \mathbf{x}_{ij}\right\}, \quad j \in s_i,\ i = 1, \ldots, m, \tag{4.5}$$

provided that $\sum_{i=1}^m \sum_{j \in s_i} w_{j|i} \mathbf{x}_{ij} \mathbf{x}'_{ij}$ is non-singular. The resulting calibration estimator of the domain total Y_i is then obtained as

$$\hat{Y}_i^{\mathrm{LCALN}} = \sum_{j \in s_i} g_{j|i} y_{ij} = \hat{Y}_i + \left(\mathbf{X} - \hat{\mathbf{X}}\right)' \hat{\mathbf{B}}_i^N, \tag{4.6}$$

where

$$\hat{\mathbf{B}}_i^N = \left(\sum_{\ell=1}^m \sum_{j \in s_\ell} w_{j|\ell} \mathbf{x}_{\ell j} \mathbf{x}'_{\ell j}\right)^{-1} \sum_{j \in s_i} w_{j|i} \mathbf{x}_{ij} y_{ij}.$$

In contrast with the GREG estimator, the correction of $\hat{Y}_i$ in $\hat{Y}_i^{\mathrm{LCALN}}$ uses the overall population total $\mathbf{X}$ and its corresponding expansion estimator.

The LCAL (or GREG) estimator (4.3) is expected to have smaller cut-off sampling bias than (4.6) because it uses auxiliary information from each particular domain i. On the other hand, for domains with small sample sizes n_i, its variance (and small sample bias) may be large since it uses only domain-specific data. The alternative calibration estimator given in (4.6) is expected to have slightly larger cut-off sampling bias because it uses only aggregated auxiliary information at the national level, but its design-variance is expected to be smaller. We now study the properties of (4.3). To this end, consider the theoretical version of LCAL estimator (4.3), given by

$$\tilde{Y}_i^{\text{LCAL}} = \hat{Y}_i + \left(\mathbf{X}_i - \hat{\mathbf{X}}_i\right)' \mathbf{B}_{iI}. \tag{4.7}$$

Here, $\mathbf{B}_{iI} = \left(\sum_{j \in U_{iI}} \mathbf{x}_{ij}\mathbf{x}'_{ij}\right)^{-1} \sum_{j \in U_{iI}} \mathbf{x}_{ij} y_{ij}$ is the census version of $\hat{\mathbf{B}}_i$ based on the set of included units from domain i. Note that the sample s_i is drawn only from U_{iI} and thus $\hat{\mathbf{B}}_i$ estimates $\mathbf{B}_{iI}$. We decompose the bias of $\hat{Y}_i^{\text{LCAL}}$ as

$$\begin{aligned} B_\pi\left(\hat{Y}_i^{\text{LCAL}}\right) &= E_\pi\left(\hat{Y}_i^{\text{LCAL}} - \tilde{Y}_i^{\text{LCAL}}\right) + B_\pi\left(\tilde{Y}_i^{\text{LCAL}}\right), \\ &= E_\pi\left\{\left(\mathbf{X}_i - \hat{\mathbf{X}}_i\right)'\left(\hat{\mathbf{B}}_i - \mathbf{B}_{iI}\right)\right\} + B_\pi\left(\tilde{Y}_i^{\text{LCAL}}\right). \end{aligned} \tag{4.8}$$

The term $E_\pi\left\{\left(\mathbf{X}_i - \hat{\mathbf{X}}_i\right)\left(\hat{\mathbf{B}}_i - \mathbf{B}_{iI}\right)\right\}/N_i$ tends to zero as $n_i \to \infty$ regardless of whether cut-off sampling is applied or not, since $\hat{\mathbf{B}}_i$ tends to $\mathbf{B}_{iI}$. However, for small n_i this term may not be negligible; that is, the LCAL estimator has small sample bias even if $U_{iE} = \varnothing$. In the absence of cut-off sampling, the bias term $B_\pi\left(\tilde{Y}_i^{\text{LCAL}}\right)$ in (4.8) is exactly equal to zero. Under cut-off sampling, we know that $E_\pi\left(\hat{Y}_i\right) = Y_{iI}$ and $E_\pi\left(\hat{\mathbf{X}}_i\right) = \mathbf{X}_{iI}$, where $\mathbf{X}_{iI} = \sum_{j \in U_{iI}} \mathbf{x}_{ij}$. Noting that $\mathbf{X}_i - \mathbf{X}_{iI} = \mathbf{X}_{iE}$, for $\mathbf{X}_{iE} = \sum_{j \in U_{iE}} \mathbf{x}_{ij}$, we obtain the design-bias of this LCAL theoretical estimator, given in absolute and relative terms by

$$B_\pi\left(\tilde{Y}_i^{\text{LCAL}}\right) = -N_{iE}\left(\bar{Y}_{iE} - \bar{\mathbf{X}}'_{iE}\mathbf{B}_{iI}\right), \quad \text{RB}_\pi\left(\tilde{Y}_i^{\text{LCAL}}\right) = -\frac{N_{iE}}{N_i}\,\frac{\bar{Y}_{iE} - \bar{\mathbf{X}}'_{iE}\mathbf{B}_{iI}}{\bar{Y}_i}. \tag{4.9}$$

This bias is small when the same model holds for the included and excluded individuals.

Since the calibration estimator $\hat{Y}_i^{\text{LCAL}}$ is intended to estimate Y_i (and not Y_{iI}), for the domain mean $\bar{Y}_i = Y_i / N_i$ we consider the estimator obtained simply dividing $\hat{Y}_i^{\text{CAL}}$ by N_i (instead of N_{iI}), $\hat{\bar{Y}}_i^{\text{LCAL}} = \hat{Y}_i^{\text{LCAL}} / N_i$. The asymptotic bias of $\hat{\bar{Y}}_i^{\text{LCAL}}$ is given by (4.9) divided by N_i.

We now analyze properties under the model and the sampling replication mechanism. Note that $\hat{\mathbf{B}}_i$ in the GREG estimator is the weighted least squares (WLS) estimator of the vector of regression coefficients $\boldsymbol{\beta}_i$ in the following linear regression model for the units in domain i:

$$y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta}_i + \varepsilon_{ij}, \; E_m\left(\varepsilon_{ij}\right) = 0, \; E_m\left(\varepsilon_{ij}^2\right) = \sigma_\varepsilon^2, \quad j = 1, \ldots, N_i, \tag{4.10}$$

where model errors ε_{ij} are all mutually independent. We wish to see the value added by the model to the design properties of the estimators; that is, how much would be gained if data were actually generated (at least approximately) by the assumed model. Let E_m denote expectation under model (4.10). If the linear regression model (4.10) actually holds for all the units in the domain (included and excluded), then $E_m(\mathbf{B}_{iI}) = \boldsymbol{\beta}_i$ and taking expectation of the bias term in (4.9) under model (4.10), we obtain the model-design bias,

$$B_{m,\pi}\left(\tilde{Y}_i^{\text{LCAL}}\right) = -N_{iE}\left\{E_m\left(\bar{Y}_{iE}\right) - \bar{\mathbf{X}}'_{iE}E_m\left(\mathbf{B}_{iI}\right)\right\} = -N_{iE}\left(\bar{\mathbf{X}}'_{iE}\boldsymbol{\beta}_i - \bar{\mathbf{X}}'_{iE}\boldsymbol{\beta}_i\right) = 0. \qquad (4.11)$$

In contrast, assuming exactly the same regression model, the bias of the basic direct estimator $\hat{Y}_i^{\text{HA}}$ under cut-off sampling is not zero unless the means of the auxiliary variables for the excluded and included units are equal. Indeed,

$$B_{m,\pi}\left(\hat{Y}_i^{\text{HA}}\right) = N_{iE}E_m\left(\bar{Y}_{iI} - \bar{Y}_{iE}\right) = N_{iE}\left(\bar{\mathbf{X}}_{iI} - \bar{\mathbf{X}}_{iE}\right)'\boldsymbol{\beta}_i. \qquad (4.12)$$

Thus, the condition under which the LCAL estimator is design-unbiased, namely that the linear model (4.10) holds without error for all the units in the domain, is much weaker than the requirements for the basic direct estimator to be design-unbiased. This means that calibration estimators will tend to be less biased than the basic direct estimator and can reduce substantially the cut-off sampling bias if the outcome is generated by the above domain-specific linear regression model.

Turning now to LCALN estimator (4.6), we define the corresponding theoretical version

$$\tilde{Y}_i^{\text{LCALN}} = \hat{Y}_i + \left(\mathbf{X} - \hat{\mathbf{X}}\right)'\mathbf{B}_{iI}^N, \qquad (4.13)$$

where $\mathbf{B}_i^N$ is the census version for the included units,

$$\mathbf{B}_i^N = \left(\sum_{\ell=1}^{m}\sum_{j\in U_{\ell I}}\mathbf{x}_{\ell j}\mathbf{x}'_{\ell j}\right)^{-1}\sum_{j\in U_{iI}}\mathbf{x}_{ij}y_{ij}.$$

Decomposing the bias similarly as in (4.8), we obtain

$$B_\pi\left(\hat{Y}_i^{\text{LCALN}}\right) = E_\pi\left\{\left(\mathbf{X} - \hat{\mathbf{X}}\right)'\left(\hat{\mathbf{B}}_i^N - \mathbf{B}_{iI}^N\right)\right\} + B_\pi\left(\tilde{Y}_i^{\text{LCALN}}\right). \qquad (4.14)$$

Again, $E_\pi\left\{\left(\mathbf{X} - \hat{\mathbf{X}}\right)\left(\hat{\mathbf{B}}_i^N - \mathbf{B}_{iI}^N\right)\right\}/N_i$ is not zero for small n_i but it tends to zero as $n_i \to \infty$ even under cut-off sampling, whereas $B_\pi\left(\tilde{Y}_i^{\text{LCALN}}\right) = 0$ only in the absence of cut-off sampling bias. In general, using the decomposition $\mathbf{X} = \mathbf{X}_I + \mathbf{X}_E$, where $\mathbf{X}_I$ and $\mathbf{X}_E$ are the national totals for the included and excluded units respectively, the design bias of $\tilde{Y}_i^{\text{LCALN}}$ is given by

$$B_\pi\left(\tilde{Y}_i^{\text{LCALN}}\right) = -\left(Y_{iE} - \mathbf{X}'_E\mathbf{B}_{iI}^N\right). \qquad (4.15)$$

Consider now the linear model with constant regression coefficients for all the population units, called model m_2:

$$y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \varepsilon_{ij},\ E_{m_2}\left(\varepsilon_{ij}\right) = 0,\ E_{m_2}\left(\varepsilon_{ij}^2\right) = \sigma_\varepsilon^2, \quad j = 1, \ldots, N_i,\ i = 1, \ldots, m, \qquad (4.16)$$

where again the model errors ε_{ij} are mutually independent. Note that, under this model, $E_{m_2}\left(\mathbf{B}_{iI}^{N}\right) \neq \boldsymbol{\beta}$ in general, but if we consider the sum $\mathbf{B}_I = \sum_{i=1}^{m} \mathbf{B}_{iI}^{N}$ instead, we have $E_{m_2}(\mathbf{B}_I) = \boldsymbol{\beta}$. This means that the theoretical LCALN estimator for a particular domain, $\tilde{Y}_i^{\text{LCALN}}$, is not model-design unbiased, because

$$B_{m_2,\pi}\left(\tilde{Y}_i^{\text{LCALN}}\right) = -\left\{\mathbf{X}'_{iE}\boldsymbol{\beta} - \mathbf{X}'_{E} E_{m_2}\left(\mathbf{B}_{iI}^{N}\right)\right\},$$

is not necessarily equal to zero. However, the national estimator obtained adding those of the domains, $\tilde{Y}^{\text{LCALN}} = \sum_{i=1}^{m} \tilde{Y}_i^{\text{LCALN}} = \hat{Y} + \left(\mathbf{X} - \hat{\mathbf{X}}\right)' \mathbf{B}_I$, is actually model-design unbiased, because

$$B_{m_2,\pi}\left(\tilde{Y}^{\text{LCALN}}\right) = -\left\{\mathbf{X}'_{E}\boldsymbol{\beta} - \mathbf{X}'_{E} E_{m_2}\left(\mathbf{B}_{I}\right)\right\} = 0.$$

Hence, under model (4.16) with constant regression coefficients for all the population units, the LCALN estimator is not model-design unbiased for a particular domain, but it is unbiased when aggregating for all the domains, provided that the same model holds for the included and excluded units in all domains. For the mean $\bar{Y}_i$, the bias of the theoretical estimator $\tilde{\bar{Y}}_i^{\text{LCALN}} = \tilde{Y}_i^{\text{LCALN}} / N_i$ is given by (4.15) divided by N_i.

We now study the variances. For the theoretical LCAL estimator (4.7), the design-variance is given by

$$V_\pi\left(\tilde{Y}_i^{\text{LCAL}}\right) = V_\pi\left(\hat{Y}_i - \hat{\mathbf{X}}'_i \mathbf{B}_{iI}\right) = V_\pi\left(\sum_{j \in s_i} w_{j|i} E_{ij}\right), \tag{4.17}$$

where $E_{ij} = y_{ij} - \mathbf{x}'_{ij}\mathbf{B}_{iI}$, $j \in U_{iI}$. We can then apply the usual variance estimators for expansion type estimators. In the case of LCALN given in (4.13), the variance is given by

$$V_\pi\left(\tilde{Y}_i^{\text{LCALN}}\right) = V_\pi\left(\hat{Y}_i - \hat{\mathbf{X}}'\mathbf{B}_{iI}^{N}\right).$$

Note that $\hat{\mathbf{X}}$ is based on the n sample units, whereas $\hat{\mathbf{X}}_i$ uses only the n_i units in domain i. As a consequence, the contribution of $\hat{\mathbf{X}}$ to the variance of LCALN should be much smaller than the contribution of $\hat{\mathbf{X}}_i$ in (4.17). This means that, provided that the domain and national regression lines are similar, the variance of LCALN estimator, obtained from the calibration at the national level, should be smaller than that of the domain-specific calibration estimator LCAL.

5 EBLUP under the nested error model

Estimators described so far use only the outcome information coming from the domain. This means that, when the domain sample size n_i is small, these estimators might be inefficient even in the absence of cut-off sampling. Small area (or indirect) estimation methods are designed to reduce the variance by increasing the effective sample size; see Rao and Molina (2015) for a comprehensive account of small area estimation methods. In this section, we focus on model-based methods, which provide estimators with good properties under the distribution induced by the model. Since the model-based properties are well known, we wish to analyze whether the estimators have good properties under the sampling-replication mechanism, which does not assume that the model actually holds.

We consider a very popular unit level model introduced by Battese, Harter and Fuller (1988) and often called nested error model. Similarly as for model m_2 in (4.16), this model assumes a constant linear regression for all the population units, but allows for unexplained heterogeneity between the domains by including random domain effects u_i apart from model errors e_{ij}. This model, denoted model m_3, assumes

$$\begin{aligned} y_{ij} &= \mathbf{x}_{ij}'\boldsymbol{\beta} + u_i + e_{ij},\ u_i \overset{\text{iid}}{\sim} N(0, \sigma_u^2), \\ e_{ij} &\overset{\text{iid}}{\sim} N(0, \sigma_e^2),\ j = 1, \ldots, N_i,\ i = 1, \ldots, m, \end{aligned} \tag{5.1}$$

where area effects u_i and errors e_{ij} are all mutually independent. The vectors $\boldsymbol{\beta}$ and $\boldsymbol{\theta} = (\sigma_u^2, \sigma_e^2)'$ are unknown. Setting $\sigma_u^2 = 0$ in (5.1), we obtain model m_2 given in (4.16). If $\mathbf{y}_i = (y_{i1}, \ldots, y_{iN_i})'$ denotes the vector of outcomes for domain i and $\mathbf{X}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iN_i})'$ the corresponding design matrix, the model in matrix notation reads

$$\mathbf{y}_i \overset{\text{ind}}{\sim} N(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{V}_i),\ \mathbf{V}_i = \sigma_u^2 \mathbf{1}_{N_i}\mathbf{1}_{N_i}' + \sigma_e^2 \mathbf{I}_{N_i},\ i = 1, \ldots, m, \tag{5.2}$$

where $\mathbf{1}_k$ denotes a vector of ones of size k and $\mathbf{I}_k$ is the $k \times k$ identity matrix.

We consider linear domain parameters defined as $H_i = \mathbf{b}_i'\mathbf{y}_i$, where $\mathbf{b}_i$ is a non-stochastic vector of known elements. The domain mean $H_i = \bar{Y}_i = N_i^{-1}\sum_{j=1}^{N_i} y_{ij}$ is obtained with $\mathbf{b}_i = N_i^{-1}\mathbf{1}_{N_i}$.

A sample s_i is supposed to be drawn from the set of included units in domain i, that is, $s_i \subset U_{iI}$. We denote by $r_i = (U_{iI} - s_i) \cup U_{iE}$ the set of non-sampled units from domain U_i, which includes those non-sampled units from U_{iI} and all the units in U_{iE}. Note that $U_i = s_i \cup r_i = U_{iI} \cup U_{iE}$. Then, the overall sample s is composed of the samples s_i drawn from the sets of included units in each area U_{iI}, $i = 1, \ldots, m$, that is, $s = s_1 \cup \cdots \cup s_m$.

We decompose the domain vector $\mathbf{y}_i$ and the design and covariance matrices $\mathbf{X}_i$ and $\mathbf{V}_i$ into the corresponding subvectors and submatrices for sample and out-of-sample units, indicated with subscripts s and r respectively, as follows

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_{is} \\ \mathbf{y}_{ir} \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{X}_{is} \\ \mathbf{X}_{ir} \end{pmatrix}, \quad \mathbf{V}_i = \begin{pmatrix} \mathbf{V}_{is} & \mathbf{V}_{isr} \\ \mathbf{V}_{irs} & \mathbf{V}_{ir} \end{pmatrix}.$$

The linear parameter $H_i = \mathbf{b}_i'\mathbf{y}_i$ can then be expressed as $H_i = \mathbf{b}_{is}'\mathbf{y}_{is} + \mathbf{b}_{ir}'\mathbf{y}_{ir}$. Under model (5.1), the best linear unbiased predictor (BLUP) of H is the model-unbiased linear function of the sample data $\hat{H}_i = \boldsymbol{\alpha}_{is}'\mathbf{y}_{is}$, which minimizes the model mean squared error (MSE), $\text{MSE}_{m_3}(\hat{H}_i) = E_{m_3}(\hat{H}_i - H_i)^2$. The BLUP of $H_i = \mathbf{b}_{is}'\mathbf{y}_{is} + \mathbf{b}_{ir}'\mathbf{y}_{ir}$ is then

$$\hat{H}_i^{\text{BLUP}}(\boldsymbol{\theta}) = \mathbf{b}_{is}'\mathbf{y}_{is} + \mathbf{b}_{ir}'\left[\mathbf{X}_{ir}\tilde{\boldsymbol{\beta}}_s + \mathbf{V}_{irs}\mathbf{V}_{is}^{-1}\left(\mathbf{y}_{is} - \mathbf{X}_{is}'\tilde{\boldsymbol{\beta}}_s\right)\right], \tag{5.3}$$

where $\tilde{\boldsymbol\beta}_s$ is the weighted least squares estimator of $\boldsymbol\beta$, given by

$$\tilde{\boldsymbol\beta}_s = \tilde{\boldsymbol\beta}_s(\boldsymbol\theta) = \left(\sum_{i=1}^{m}\mathbf{X}'_{is}\mathbf{V}_{is}^{-1}\mathbf{X}_{is}\right)^{-1}\sum_{i=1}^{m}\mathbf{X}'_{is}\mathbf{V}_{is}^{-1}\mathbf{y}_{is}. \tag{5.4}$$

The BLUP of H_i given in (5.3) depends on the true values of the variance components $\boldsymbol\theta = (\sigma_u^2, \sigma_e^2)'$, which are typically unknown. Replacing them by corresponding model-consistent estimators $\hat{\boldsymbol\theta} = (\hat\sigma_u^2, \hat\sigma_e^2)'$, we obtain the so-called empirical BLUP (EBLUP), denoted $\hat{H}_i^{\text{EBLUP}} = \hat{H}_i^{\text{BLUP}}(\hat{\boldsymbol\theta})$.

If the domain sampling fraction, n_i/N_i, is negligible, the BLUP of $\bar{Y}_i$ may be expressed as the weighted average

$$\hat{\bar{Y}}_i^{\text{BLUP}} \cong \gamma_{is}\left[\bar{y}_{is} + (\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_{is})'\,\tilde{\boldsymbol\beta}_s\right] + (1-\gamma_{is})\,\bar{\mathbf{X}}_i'\tilde{\boldsymbol\beta}_s, \tag{5.5}$$

where $\gamma_{is} = \sigma_u^2/(\sigma_u^2 + \sigma_e^2/n_i)$ is in the $(0, 1)$ interval and tends to 1 as $n_i \to \infty$ (Rao and Molina, 2015). Thus, for domains with large sample size n_i, $\hat{\bar{Y}}_i^{\text{BLUP}}$ approaches the survey regression estimator $\bar{y}_{is} + (\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_{is})'\,\tilde{\boldsymbol\beta}_s$, whereas for domains with small sample size n_i, $\hat{\bar{Y}}_i^{\text{BLUP}}$ borrows strength from the other domains by approaching the regression-synthetic estimator $\bar{\mathbf{X}}_i'\tilde{\boldsymbol\beta}_s$. Replacing the variance components in $\boldsymbol\theta = (\sigma_u^2, \sigma_e^2)'$ by consistent estimators $\hat{\boldsymbol\theta} = (\hat\sigma_u^2, \hat\sigma_e^2)'$ in the BLUP, denoting $\hat\gamma_{is} = \hat\sigma_u^2/(\hat\sigma_u^2 + \hat\sigma_e^2/n_i)$ and $\hat{\boldsymbol\beta}_s = \tilde{\boldsymbol\beta}_s(\hat{\boldsymbol\theta})$, we obtain the EBLUP of $\bar{Y}_i$, given by

$$\hat{\bar{Y}}_i^{\text{EBLUP}} \cong \hat\gamma_{is}\left[\bar{y}_{is} + (\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_{is})'\,\hat{\boldsymbol\beta}_s\right] + (1-\hat\gamma_{is})\,\bar{\mathbf{X}}_i'\hat{\boldsymbol\beta}_s. \tag{5.6}$$

The BLUP is unbiased and optimal under model m_3 in the sense of minimizing the MSE under that model. We now study its design properties, which do not assume that the model is correct and hence account for bias under model departures. To that end, we consider the census regression parameter for the included units, defined as $\mathbf{B}_I = \left(\sum_{i=1}^m \mathbf{X}'_{iI}\mathbf{V}_{iI}^{-1}\mathbf{X}_{iI}\right)^{-1}\sum_{i=1}^m \mathbf{X}'_{iI}\mathbf{V}_{iI}^{-1}\mathbf{y}_{iI}$, where $\mathbf{y}_{iI}$, $\mathbf{X}_{iI}$ and $\mathbf{V}_{iI}$ are the corresponding sub-vector and sub-matrices of $\mathbf{y}_i$, $\mathbf{X}_i$ and $\mathbf{V}_i$, for the included units $(j \in U_{iI})$. Again, we consider the theoretical version of the BLUP defined in terms of $\mathbf{B}_I$,

$$\tilde{\bar{Y}}_i^{\text{BLUP}} = \gamma_{is}\left\lfloor\bar{y}_{is} + (\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_{is})'\mathbf{B}_I\right\rfloor + (1-\gamma_{is})\,\bar{\mathbf{X}}_i'\mathbf{B}_I. \tag{5.7}$$

If each sample s_i is drawn from the corresponding domain U_{iI} by simple random sampling without replacement (SRSWOR), then $E_\pi(\bar{y}_{is}) = \bar{Y}_{iI}$ and $E_\pi(\bar{\mathbf{x}}_{is}) = \bar{\mathbf{X}}_{iI}$. Using these facts, it is easy to calculate the design-bias of $\tilde{\bar{Y}}_i^{\text{BLUP}}$ under SRSWOR, which is given by

$$B_\pi\left(\tilde{\bar{Y}}_i^{\text{BLUP}}\right) = \gamma_{is}\frac{N_{iE}}{N_{iI}}\left[\left(\bar{Y}_i - \bar{\mathbf{X}}_i'\mathbf{B}_I\right) - \left(\bar{Y}_{iE} - \bar{\mathbf{X}}_{iE}'\mathbf{B}_I\right)\right] + (1-\gamma_{is})\left(\bar{\mathbf{X}}_i'\mathbf{B}_I - \bar{Y}_i\right). \tag{5.8}$$

This bias will be small if (5.1) holds for the whole population, in which case $E_{m_3}(\bar{Y}_i) = \bar{\mathbf{X}}_i\boldsymbol\beta$ and $E_{m_3}(\bar{Y}_{iE}) = \bar{\mathbf{X}}_{iE}\boldsymbol\beta$. Using these results when taking expectation under model m_3 in (5.8), we get $B_{m_3,\pi}\left(\tilde{\bar{Y}}_i^{\text{BLUP}}\right) = 0$. In fact, the same result also holds under model m_2.

Concerning variance, if s_i is obtained by SRSWOR within U_{iI}, the design-variance of the theoretical BLUP estimator is given by

$$V_\pi\left(\tilde{\bar{Y}}_i^{\text{BLUP}}\right) = \gamma_{is}^2 V_\pi\left(\bar{y}_{is} - \bar{\mathbf{x}}_{is}\mathbf{B}_I\right) = \frac{\gamma_{is}^2}{N_i^2} V_\pi\left(\hat{Y}_i - \hat{\mathbf{X}}_i'\mathbf{B}_I\right).$$

Hence, if the census least squared (LS) regression lines for the domains from model (4.10) are similar to the national census weighted least squared (WLS) regression line from model (5.1), that is, if $\mathbf{B}_I \approx \mathbf{B}_{iI}$, then the variance of the BLUP for $\bar{Y}_i$ reduces to that of the LCAL estimator of $\bar{Y}_i$ obtained from (4.17), multiplied by the factor $\gamma_{is}^2 \in (0, 1)$.

Under more general sampling designs within U_{iI}, we consider the pseudo-EBLUP of $\bar{Y}_i$ proposed by You and Rao (2002) instead of the EBLUP. Defining the analogous theoretical estimator that uses the weighted sample means $\bar{y}_{iw} = \left(\sum_{j \in s_i} w_{j|i}\right)^{-1} \sum_{j \in s_i} w_{j|i} y_{ij}$ and $\bar{\mathbf{x}}_{iw} = \left(\sum_{j \in s_i} w_{j|i}\right)^{-1} \sum_{j \in s_i} w_{j|i} \mathbf{x}_{ij}$ instead or the unweighted ones $\bar{y}_{is}$ and $\bar{\mathbf{x}}_{is}$ in (5.7), we obtain the same expressions for the design bias and variance, with γ_{is} changed to $\gamma_{iw} = \sigma_u^2 / \left(\sigma_u^2 + \sigma_e^2 \delta_{iw}\right)$, for $\delta_{iw} = \left(\sum_{j \in s_i} w_{j|i}\right)^{-2} \sum_{j \in s_i} w_{j|i}^2$.

6 Empirical best predictor under the nested error model

Estimation of non-linear domain parameters requires more general small area estimation methods, such as the best/Bayes predictor (BP), see Molina and Rao (2010). Special non-linear parameters are poverty and inequality indicators defined in terms of a welfare measure, such as the family of poverty indicators introduced by Foster, Greer and Thorbecke (1984). The best predictor can also be used for the estimation of other characteristics such as median, quantiles or even the whole empirical distribution function of the variable of interest, see Pratesi (2016). Additionally, it can be used to estimate totals and means of a given target variable, when the dependent variable in the considered model is a one-to-one transformation (e.g., log or more general Box-Cox transformations) of this target variable. These transformations are typically applied in the case of non-normality or heteroscedasticity.

In this section, the target variable (e.g., the welfare measure) for the j^{th} unit in i^{th} domain is denoted as v_{ij} and $y_{ij} = T\left(v_{ij}\right)$, where T is a one-to-one transformation. We assume that y_{ij} follows the nested error model (5.1). By the inverse transformation $v_{ij} = T^{-1}\left(y_{ij}\right)$, we can express our target parameter (defined originally in terms of the target variables v_{ij}) as a function of the vector $\mathbf{y}_i = \left(y_{i1}, \ldots, y_{iN_i}\right)'$ of model responses for the domain units, $H_i = h\left(\mathbf{y}_i\right)$. The best predictor (BP) of $H_i = h\left(\mathbf{y}_i\right)$ is defined as the function of the sample data $\mathbf{y}_{is}$ that minimizes the model MSE, and it turns out to be

$$\hat{H}_i^{\text{BP}}(\boldsymbol{\beta}, \boldsymbol{\theta}) = E_{m_3}\left[h(\mathbf{y}_i) | \mathbf{y}_{is}; \boldsymbol{\beta}, \boldsymbol{\theta}\right], \tag{6.1}$$

where the expectation is taken with respect to the model distribution of $\mathbf{y}_{ir} | \mathbf{y}_{is}$, which depends on the true values of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. The BP of H_i is unbiased with respect to the model (5.1), regardless of the complexity of the function $h(\cdot)$ defining the target parameter. However, it cannot be calculated in practice since model parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are typically unknown. An empirical best predictor (EBP) of H_i, denoted as $\hat{H}_i^{\text{EBP}}$, is then obtained by replacing $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ in $\hat{H}_i^{\text{BP}}(\boldsymbol{\beta}, \boldsymbol{\theta})$ by consistent estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$, as $\hat{H}_i^{\text{EBP}} = \hat{H}_i^{\text{BP}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$. The EBP is not exactly unbiased, but the bias arising from the estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ is typically negligible when the overall sample size n is large. In the case of a linear parameter $H_i = \mathbf{b}_i'\mathbf{y}_i$, the EBP under the nested error model with normality obtained using $\hat{\boldsymbol{\beta}}_s = \tilde{\boldsymbol{\beta}}_s(\hat{\boldsymbol{\theta}})$ to estimate $\boldsymbol{\beta}$ equals $\hat{H}_i^{\text{EBLUP}}$.

When $h(\cdot)$ is so complex that the expectation defining the EBP in (6.1) cannot be calculated analytically, Monte Carlo methods can be applied to approximate $\hat{H}_i^{\text{EBP}}$ as proposed in Molina and Rao (2010). This is done by simulating, from the model (5.1) fitted to the original sample data, L replicates $y_{ij}^{(\ell)}$; $\ell = 1, \ldots, L$ of y_{ij}, $j \in r_i$, where r_i are the non-sample units of area i, attaching the sample elements y_{ij}, $j \in s_i$ to form the population vector $\mathbf{y}_i^{(\ell)}$, calculating the corresponding target parameter $H_i^{(\ell)} = h(\mathbf{y}_i^{(\ell)})$ for each $\ell = 1, \ldots, L$ and, finally, averaging over the L replicates as $\hat{H}_i^{\text{EBP}} = L^{-1}\sum_{\ell=1}^{L} H_i^{(\ell)}$. Note that the EBP requires the values $\mathbf{x}_{ij}$ for all units in the population, and not only for the included units. For further details, see Molina and Rao (2010).

7 MSE estimation

The EBLUP in Section 5 or the EBP described in Section 6 are based on the nested error model (5.1). Calibration estimators described in Section 4 are also assisted by a linear regression model. If we wish to have comparable accuracy measures, it seems reasonable to obtain the MSEs of all the estimators under a given regression model (model MSE), assuming that the model holds for all the population units (included and excluded). Here, we estimate the model MSE using the bootstrap method proposed in Molina and Rao (2010), which is based on the original parametric bootstrap method for finite populations of González-Manteiga, Lombardia, Molina, Morales and Santamaría (2008). According to this procedure, the bootstrap MSE of $\hat{H}_i^{\text{EBP}}$ under the nested error model (5.1) is obtained as follows: i) Fit Model (5.1) to the sample data $\{(\mathbf{y}_{is}, \mathbf{X}_{is}); i = 1, \ldots, m\}$, to obtain the estimators $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}_u^2$ and $\hat{\sigma}_e^2$ of $\boldsymbol{\beta}$, σ_u^2 and σ_e^2 respectively. ii) For $b = 1, \ldots, B$, generate independently $u_i^{*(b)} \overset{\text{iid}}{\sim} N(0, \hat{\sigma}_u^2)$ and $e_{ij}^{*(b)} \overset{\text{iid}}{\sim} N(0, \hat{\sigma}_e^2)$, $j = 1, \ldots, N_i$, $i = 1, \ldots, m$. iii) For $b = 1, \ldots, B$, construct bootstrap domain vectors $\mathbf{y}_i^{*(b)} = (y_{i1}^{*(b)}, \ldots, y_{iN_1}^{*(b)})'$, whose elements are generated as

$$y_{ij}^{*(b)} = \mathbf{x}_{ij}'\hat{\boldsymbol{\beta}} + u_i^{*(b)} + e_{ij}^{*(b)}, \quad j = 1, \ldots, N_i, i = 1, \ldots, m.$$

From the bootstrap domain vector $\mathbf{y}_i^{*(b)}$, calculate the target bootstrap parameter $H_i^{*(b)} = h(\mathbf{y}_i^{*(b)})$, for $b = 1, \ldots, B$. iv) From each bootstrap population vector $\mathbf{y}_i^{*(b)}$, take the sample part $\mathbf{y}_{is}^{*(b)}$, where the sample indices s_i are exactly those of the original sample drawn from U_{iI}, for $i = 1, \ldots, m$. Using the overall bootstrap sample data $\mathbf{y}_s^{*(b)} = (\mathbf{y}_{1s}^{*(b)}, \ldots, \mathbf{y}_{ms}^{*(b)})'$ and the population vectors $\mathbf{x}_{ij}$, $j = 1, \ldots, N_i$, assumed to be known for all population units, calculate the bootstrap EBP of H_i, denoted as $\hat{H}_i^{\text{EBP}*(b)}$, $b = 1, \ldots, B$. v) A bootstrap MSE estimator for the EBP under model (5.1), $\text{MSE}_{m_3}(\hat{H}_i^{\text{EBP}})$, is obtained as

$$\operatorname{mse}_B\left(\hat{H}_i^{\mathrm{EBP}}\right) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{H}_i^{\mathrm{EBP}*(b)} - H_i^{*(b)}\right)^2. \tag{7.1}$$

Bootstrap estimators of the MSE under the same model of the calibration estimators can be obtained similarly. For the special case of a linear parameter, $H_i = \mathbf{b}_i'\mathbf{y}_i$, if $\hat{\boldsymbol{\beta}}_s$ is the WLS estimator (5.4), then (7.1) is actually an estimator of $\mathrm{MSE}_{m_3}\left(\hat{H}_i^{\mathrm{EBLUP}}\right)$. This naïve bootstrap estimator of the model MSE is first-order unbiased in the sense that its model bias is $O\left(m^{-1}\right)$, but not $o\left(m^{-1}\right)$. Bias corrections existing in the literature increase the variance and may yield negative MSE estimates. In the literature, we cannot find bootstrap estimators of the MSE that are strictly positive and also second-order unbiased. Thus, for simplicity, we consider the naive bootstrap estimator (7.1), which cannot yield negative values and performs well for moderate number of areas m.

8 Simulation experiments

8.1 Aims and general description

In this section, we describe simulation experiments designed to compare the small sample properties of the estimators of $\bar{Y}_i$ discussed above in the context of cut-off sampling. Specifically, we compare the naïve direct estimator $\hat{\bar{Y}}_i^{\mathrm{HA}}$, calibration estimators $\hat{\bar{Y}}_i^{\mathrm{LCAL}}$ and $\hat{\bar{Y}}_i^{\mathrm{LCALN}}$, and the EBLUP under the nested error model $\hat{\bar{Y}}_i^{\mathrm{EBLUP}}$, under two different scenarios. In the first scenario, the values of the target variable for all the population units are generated from the same model; in the second, included and excluded units are generated from different models.

In the absence of cut-off sampling, calibration estimators are design-consistent as the domain size n_i increases even if the corresponding model does not hold, but this is not the case for model-based estimators. On the other hand, under the corresponding model, the EBLUP of a linear parameter is approximately the most efficient linear and unbiased estimator, so making simulations under a model would not provide any additional knowledge. The purpose here is to see whether the model-based predictors also perform well with respect to the (cut-off sampling) design. For this reason, we run design-based simulations by generating one population vector $\mathbf{y} = \left(\mathbf{y}_1', \ldots, \mathbf{y}_m'\right)'$ from the nested error model in (5.1), keeping it fixed and repeatedly drawing a new cut-off sample in each MC simulation. Allocation of units to the sets of included or excluded units is done by generating a random binary variable c_{ij} for each unit $j = 1, \ldots, N_i$ and area $i = 1, \ldots, m$. The units j with $c_{ij} = 1$ are assigned to U_{iI} and those with $c_{ij} = 0$ to U_{iE}. In each Monte Carlo (MC) replicate, samples are drawn, independently for each domain i, from the U_{iI} units, $i = 1, \ldots, m$.

8.2 Common regression model

We consider a population of $N = 20{,}000$ individuals divided into $m = 80$ domains with the same size $N_i = 250$, $i = 1, \ldots, m$. We consider three auxiliary variables, with values generated as $x_{ij\kappa} \overset{\text{iid}}{\sim} N(3, 2)$, $\kappa = 1, 2, 3$. The binary variables c_{ij} determining the allocation of units in U_{iI} or U_{iE} for each domain i

are generated independently as $c_{ij} \overset{\text{ind}}{\sim} \text{Bern}\left(p_{j|i}\right)$, where the probabilities $p_{j|i} = \Pr\left(c_{ij} = 1\right)$ are related to the vector of auxiliary variables $\mathbf{x}_{ij} = \left(x_{ij1}, x_{ij2}, x_{ij3}\right)'$ in the form

$$p_{j|i} = \frac{\exp\left(\mathbf{x}'_{ij}\boldsymbol{\zeta}\right)}{1 + \exp\left(\mathbf{x}'_{ij}\boldsymbol{\zeta}\right)}, \quad j = 1, \ldots, N_i, \ i = 1, \ldots, m.$$

We take $\boldsymbol{\zeta} = (0.75, 1, 1)'$. Based on this value, the total number of included units (with $c_{ij} = 1$) from all the domains represents roughly half of the population.

The values of the target variable y_{ij} are generated from the nested error model (5.1) using $\mathbf{x}_{ij} = \left(x_{ij1}, x_{ij2}, x_{ij3}\right)'$ and taking $\boldsymbol{\beta} = (1, 1.5, 1)'$, $\sigma_u^2 = (0.75)^2$ and $\sigma_e^2 = 4^2$, which leads to a determination coefficient $R^2 \approx 0.5$. Then, keeping the population values $\{(\mathbf{x}_{ij}, y_{ij}, c_{ij});\ j = 1, \ldots, N_i,$ $i = 1, \ldots, m\}$ fixed, we draw $K = 1{,}000$ Monte Carlo samples $s^{(k)}$, $k = 1, \ldots, K$. Each of these samples is obtained by drawing independent domain sub-samples $s_i^{(k)}$ of size n_i from the units in U_{il} by SRSWOR, $i = 1, \ldots, m$. The domain sample sizes are taken as $n_i \in \{5, 10, 30, 50\}$, with each sample size repeated for 20 subsequent domains. With the data from the k^{th} sample, we compute the basic direct estimator, calibration estimators at the domain level (LCAL) and at the population level (LCALN), and EBLUP. Weights, $h_{j|i}$ and $g_{j|i}$, in the calibration estimators (4.3) and (4.6) respectively are obtained using the function calib from package `sampling` (Tillé and Matei, 2016) of R (R Development Core Team, 2016). EBLUPs are obtained using R package `sae` (Molina and Marhuenda, 2015), which by default estimates the model parameters σ_u^2, σ_e^2 and $\boldsymbol{\beta}$ using restricted maximum likelihood (REML).

Let $\hat{\bar{Y}}_i$ be a generic estimator of $\bar{Y}_i$ and $\hat{\bar{Y}}_i^{(k)}$ its value obtained with k^{th} sample. We evaluate the performance of estimators in terms of relative bias (RB) and relative root MSE (RRMSE) under the design, approximated empirically as

$$\text{RB}_\pi\left(\hat{\bar{Y}}_i\right) = 100\,\frac{K^{-1}\sum_{k=1}^{K}\left(\hat{\bar{Y}}_i^{(k)} - \bar{Y}_i\right)}{\bar{Y}_i}, \quad \text{RRMSE}_\pi\left(\hat{\bar{Y}}_i\right) = 100\,\frac{\sqrt{K^{-1}\sum_{k=1}^{K}\left(\hat{\bar{Y}}_i^{(k)} - \bar{Y}_i\right)^2}}{\bar{Y}_i}.$$

Averages across domains of absolute RB and of RRMSE are also calculated as

$$\overline{\text{ARB}} = m^{-1}\sum_{i=1}^{m}\left|\text{RB}_\pi\left(\hat{\bar{Y}}_i\right)\right|, \quad \overline{\text{RRMSE}} = m^{-1}\sum_{i=1}^{m}\text{RRMSE}_\pi\left(\hat{\bar{Y}}_i\right).$$

Figure 8.1 displays boxplots of percent RB for the considered estimators of the mean $\bar{Y}_i$, where each boxplot is for the 20 domains in each group of sample sizes $n_i = 5, 10, 30, 50$. We can see the large cut-off sampling bias of the basic direct estimator, with median RB exceeding 20% for all the domain sample sizes. This cut-off sampling bias is corrected by all the other estimators. Nevertheless, the LCALN estimator shows wider boxplots. This estimator gets large bias for some domains probably because its assisting model is not accounting for the domain effects. The LCAL estimator is based on a model that accounts for domain effects and performs well in terms of design bias uniformly for all the domain sample sizes, although EBLUP also performs rather well in terms of design bias.

Looking now at the RRMSE in Figure 8.2, we can see the much smaller RRMSEs of EBLUPs for all the domain sample sizes. The LCAL estimator gets closer RRMSEs as the domain sample size grows, but for $n_i = 5$ it gets huge RRMSEs. We have seen that the LCALN can be substantially biased for some domains and it also has large RRMSEs for all the domain sample sizes. Thus, in summary, EBLUP exhibits the lowest design RRMSE and at the same time keeps the design bias under control.

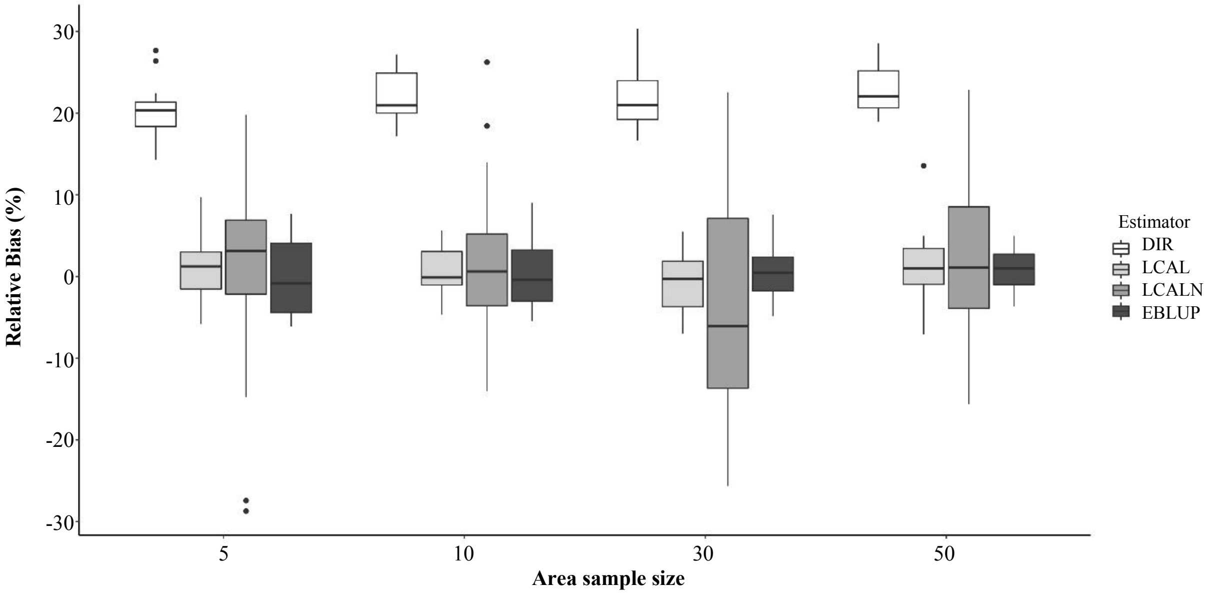

Figure 8.1 Boxplots of domain RBs (%) of basic direct, LCAL, LCALN and EBLUP estimators for n_i = 5, 10, 30, 50.

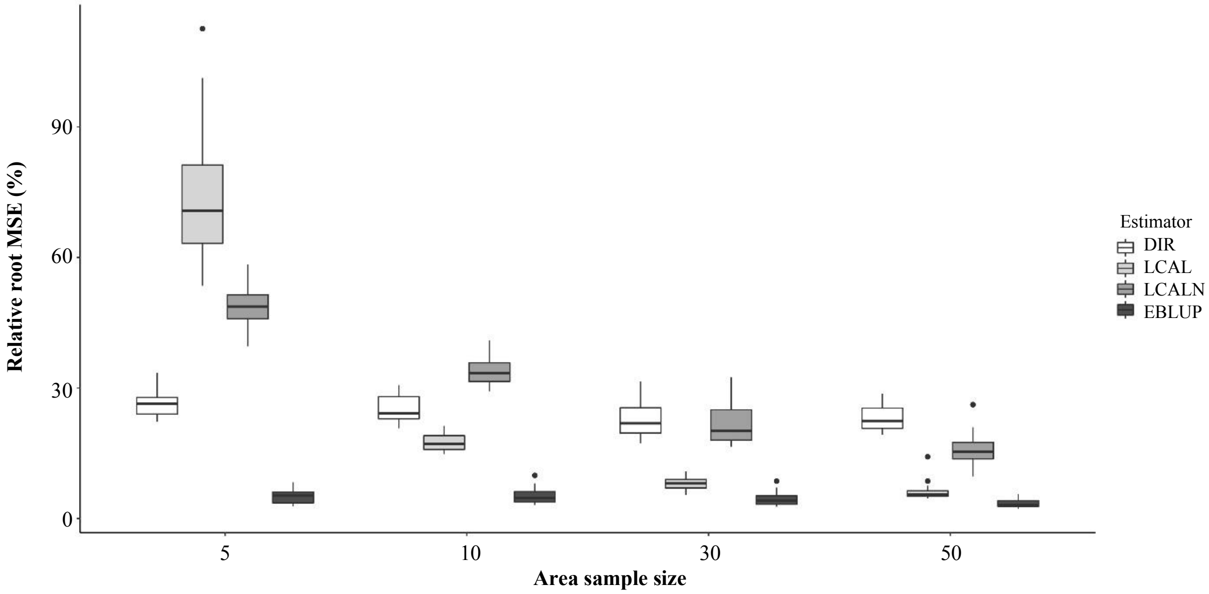

Figure 8.2 Boxplots of domain RRMSEs (%) of basic direct, LCAL, LCALN and EBLUP estimators for n_i = 5, 10, 30, 50.

Table 8.1 reports averages across all the domains of absolute RB and RRMSE, together with % share of squared bias from the total design MSE. We can see again the large cut-off sampling bias of the basic direct estimator, with a bias share of $B_\pi^2 / \mathrm{MSE}_\pi \approx 100\%$, in contrast to all other estimators. The LCAL estimator has the smallest average ARB, followed closely by EBLUP. LCALN performs the best in terms of bias ratio because of its large MSE. Thus, we consider that LCAL performs better. As already said, EBLUP clearly performs the best when looking at both bias and MSE.

Table 8.1
Averages across areas of absolute RB, RRMSE and $B_\pi^2 / \mathrm{MSE}_\pi$ for basic direct, LCAL, LCALN and EBLUP (in percentage)

Method	$\overline{\text{ARB}}$	$\overline{\text{RRMSE}}$	$B_\pi^2 / \mathrm{MSE}_\pi$
DIR	21.82	24.45	98.32
LCAL	2.96	27.33	2.48
LCALN	8.97	30.44	0.04
EBLUP	3.13	4.56	0.18

8.3 Different regression models

In this simulation experiment, we preserve the same population values and sampling scheme as before, but the values of the target variable for the included and excluded units are generated from models with different parameter values. Of course, this is not a favorable scenario for the considered model-based estimators, but it may be realistic since, in practice, the assumed model cannot be checked for the excluded units. Thus, instead of a constant $\boldsymbol{\beta}$ for all the population units, we take $\boldsymbol{\beta}_I = (1, 1.5, 1)'$ for the included units and $\boldsymbol{\beta}_E = (0.5, 1.6, 0.5)'$ for the excluded ones. The values of the explanatory variables and variance components σ_u^2 and σ_e^2 are taken exactly as before. Again, we draw $K = 1{,}000$ samples $s^{(k)}$ by independent SRSWOR within the units in domain i with $c_{ij} = 1,$ with the same domain sample sizes n_i as before. With the sample data from the k^{th} sample, we compute basic direct, LCAL, LCALN and EBLUP estimates of $\bar{Y}_i$.

Figure 8.3 shows boxplots of the corresponding percent RBs for each domain sample size. In this case, all the estimators are biased, but the bias of the basic direct estimator becomes huge, exceeding 40% for some of the domains. The bias of LCAL and EBLUP is kept relatively small for all the domains, but that of LCALN estimator is still very large in absolute value for some of the domains. In absence of cut-off sampling, the calibration estimators are asymptotically design-unbiased as the domain sample size n_i increases, even if the considered model does not hold. However, this is not true under cut-off sampling and for this reason the RBs of calibration estimators do not decrease as n_i grows. Even under this unfavorable scenario of different generating models for included and excluded units, EBLUP shows a moderate bias, which is comparable to that of LCAL estimator, and performs clearly the best in terms of RRMSE.

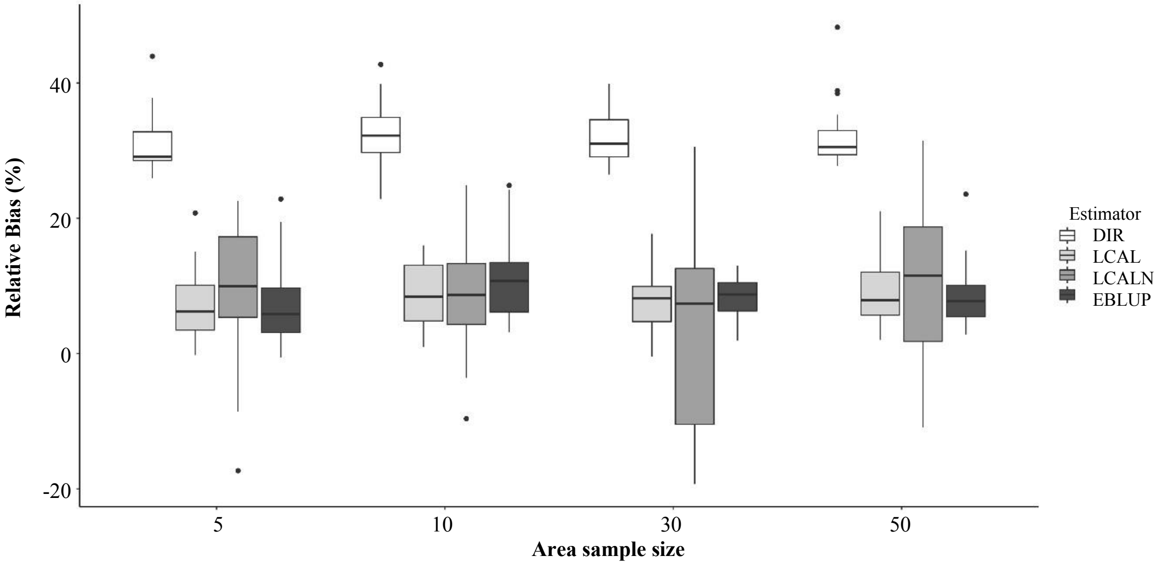

Figure 8.3 Boxplots of domain RBs (%) of basic direct, LCAL, LCALN and EBLUP estimators for n_i = 5, 10, 30, 50, when $\beta_I = (1, 1.5, 1)'$ for included units and $\beta_E = (0.5, 1.6, 0.5)'$ for excluded ones.

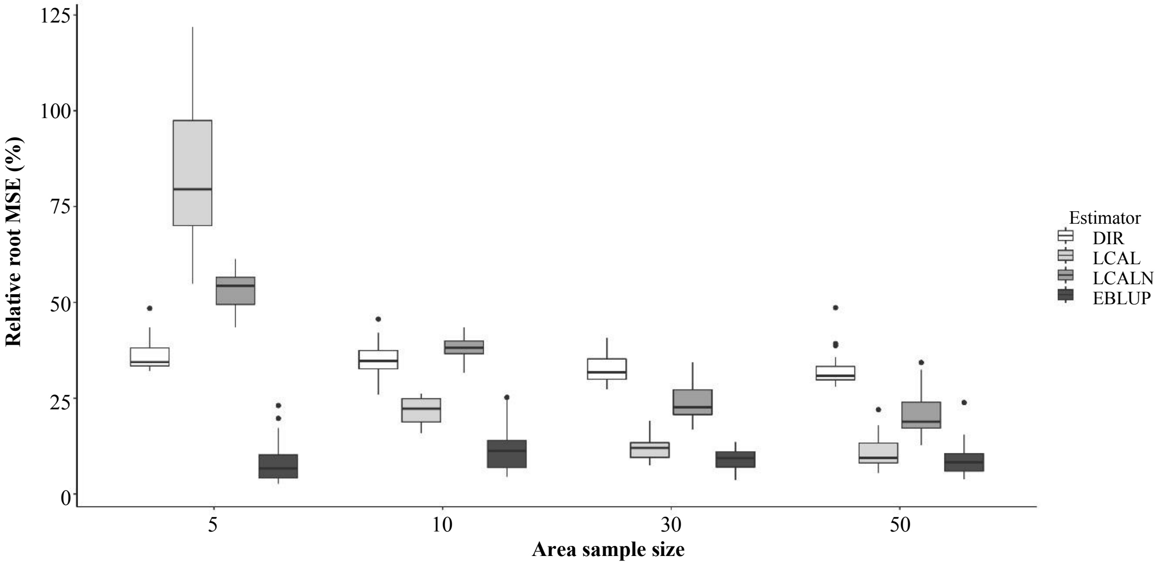

Figure 8.4 Boxplots of domain RRMSEs (%) of basic direct, LCAL, LCALN and EBLUP estimators for n_i = 5, 10, 30, 50, when $\beta_I = (1, 1.5, 1)'$ for included units and $\beta_E = (0.5, 1.6, 0.5)'$ for excluded ones.

Again, averages across all the domains of absolute RB and RRMSE are shown in Table 8.2, together with sq. bias ratio. As already noted, the basic direct estimator has a huge bias, whereas LCAL and EBLUP estimators keep an $\overline{\text{ARB}}$ below 10%. LCALN displays the lowest bias ratio because of a larger MSE. Again, EBLUP shows the best performance in terms of efficiency, with an average RRMSE also below 10%.

Table 8.2
Averages across areas of absolute RB, RRMSE and B_π^2 / MSE_π for basic direct, LCAL, LCALN and EBLUP, when $\boldsymbol{\beta}_I$ = (1, 1.5, 1)′ for included units and $\boldsymbol{\beta}_E$ = (0.5, 1.6, 0.5)′ for excluded ones (in percentage)

Method	$\overline{\text{ARB}}$	$\overline{\text{RRMSE}}$	B_π^2 / MSE_π
DIR	31.78	34.11	99.87
LCAL	8.47	30.83	77.43
LCALN	12.75	34.49	29.56
EBLUP	8.73	9.48	75.78

The simulation experiment was repeated taking a value of $\boldsymbol{\beta}_E$ further away from $\boldsymbol{\beta}_I$, making the two regression models differ substantially. Results are not included due to space constraints but, as one would expect, RB and RRMSE values increase for all estimators, but conclusions are similar to the last experiment. The basic direct estimator gets the largest RB, calibration estimators and EBLUP clearly reduce the cut-off sampling bias of the basic direct estimator and EBLUP gets smaller RRMSE, specially for the domains with the smallest sample sizes.

9 Estimation of total sales in Spanish provinces

Here we describe an application to the estimation of the total sales of a certain tobacco product in the Spanish provinces. The available data set contains, for N = 12,791 tobacco establishments (practically all of them) in m = 48 provinces from Spain (the Canary Islands, Ceuta and Melilla are not included), the volume of purchases made by each establishment of this product during the three months previous to November 2016 (z_{ij}, in euros). It also contains a variable indicating whether the establishment is supplied with a device recording all the required information about each sale. Only the establishments with larger sales are supplied with such a device. Those establishments (in total n = 1,842) are able to report proper data on sales and therefore the volume of sales (v_{ij}, in euros) of the considered product in November 2016 is also included in the data for those establishments.

We estimate the total sales $V_i = \sum_{i=1}^{N_i} v_{ij}$ in each of the m = 48 provinces included in the data using the basic direct, the selected calibration estimators and a model-based estimator. Establishments j with both z_{ij} and v_{ij} available for a province i compose the set of included units U_{iI}, which equals the sample s_i in this case (there is no sampling within U_{iI}). Then, here the basic direct estimators are given by $\hat{V}_i^{\mathrm{HA}} = N_i \bar{V}_{iI}$, $i = 1, \ldots, m$, which have actually zero variance, but might be severely biased. Since true values in real applications are not available and therefore real biases cannot be evaluated (there is no information from U_{iE}), here we will compare the estimators considering the set of establishments with sales recorded from each province as a SRSWOR from that province. Note that this is the best scenario for the basic direct estimator. Thus, for the basic direct estimator $\hat{V}_i^{\mathrm{HA}}$ considering that the actual sample $s_i = U_{iI}$ is a SRSWOR from U_i, the variance equals the MSE (we ignore the bias). A design-unbiased estimator of the MSE is then

$$\mathrm{mse}_\pi\left(\hat{V}_i\right) = N_i^2 \frac{s_i^2}{n_i}\left(1 - \frac{n_i}{N_i}\right), \quad i = 1, \ldots, m,$$

where $s_i^2 = (n_i - 1)^{-1} \sum_{j \in s_i} \left(v_{ij} - \bar{v}_{is}\right)^2$ is the sample variance of the sales for province i and here $n_i = N_{iI}$, $i = 1, \ldots, m$.

For the estimators that consider a regression model, we first make a preliminary descriptive analysis of the variables. Histograms of sales v_{ij} and of purchases z_{ij} show right-skewed distributions for both variables. Moreover, a scatterplot of ordinary LS residuals from a linear model for v_{ij} in terms of z_{ij}, against z_{ij} reveals a mild pattern of heteroscedasticity. Transforming the sales with the squared root, that is, taking $y_{ij} = v_{ij}^{1/2}$ as response variable and $\mathbf{x}_{ij} = \left(1, x_{ij}\right)'$, with $x_{ij} = z_{ij}^{1/2}$ as covariate seems to minimize the problem. Accordingly, we will consider a nested error model (5.1) for the transformed sales y_{ij} in terms of the transformed purchases x_{ij}, and EBPs of the total sales in each province, $V_i = \sum_{j=1}^{N_i} v_{ij}$, will be computed based on this model. Note that, in terms of the model responses y_{ij}, the total sales are given by $V_i = \sum_{j=1}^{N_i} y_{ij}^2 = h\left(\mathbf{y}_i\right)$. Then, the EBP of $V_i = h\left(\mathbf{y}_i\right)$ is given by $\hat{V}_i^{\mathrm{EBP}} = E_{m_3}\left[h\left(\mathbf{y}_i\right) | \mathbf{y}_{is}; \hat{\boldsymbol{\theta}}\right]$, $i = 1, \ldots, m$, which can be calculated analytically or approximated by Monte Carlo simulation. We estimate the model MSE of the EBP using the parametric bootstrap described in Section 7 for $H_i = V_i$, taking $H_i^{*(b)} = V_i^{*(b)}$ and $\hat{H}_i^{\mathrm{EBP}*(b)} = \hat{V}_i^{\mathrm{EBP}*(b)}$ and considering that the model holds for included and excluded units. Residuals from this model are described below.

Note that the LCAL (or GREG) estimator is not defined for a non-linear function of the values of the response variable in the population units, such as the total sales $V_i = \sum_{j=1}^{N_i} y_{ij}^2$ after the square root transformation. Hence, here we calculate the GREG according to (4.3) using v_{ij} instead of y_{ij} and z_{ij} instead of x_{ij}, which is assisted by the linear model (4.10) for the untransformed sales v_{ij} in terms of purchases z_{ij}. As a measure of uncertainty of the GREG, to make it comparable with that of the EBP, we estimated its model MSE through the same bootstrap procedure, replacing $\hat{H}_i^{\mathrm{EBP}*(b)}$ by $\hat{V}_i^{\mathrm{GREG}*(b)}$. The obtained bootstrap MSE estimator actually includes the error due to the fact that the correct model is the one with transformed variables.

Before comparing the estimates, we analyze the residuals from the nested error model (5.1), given by $\hat{e}_{ij} = y_{ij} - \mathbf{x}_{ij}'\hat{\boldsymbol{\beta}} - \hat{u}_i$. Figure 9.1 shows a scatterplot of those residuals against predicted values $\hat{y}_{ij} = \mathbf{x}_{ij}'\hat{\boldsymbol{\beta}} + \hat{u}_i$ (left) and a histogram of residuals (right). We can see a few negative outliers on the left plot, which agrees with a slightly larger left tail in the histogram. Apart from that, the residuals do not exhibit any remarkable pattern. In fact, in the histogram they appear to be very much concentrated around zero, which indicates a high predictive power of the model.

Figure 9.2 shows the normal Q-Q plot of predicted area effects $\hat{u}_i$. This plot supports the normality of $\hat{u}_i$ except for one outlier appearing at the left tail of the distribution. This point corresponds to the province with the smallest sample size (n_i = 3 observations), which suggests that the estimated random effect for that province, $\hat{u}_i$, is not very reliable. Thus, we consider that the nested error model fits reasonably well the available data.

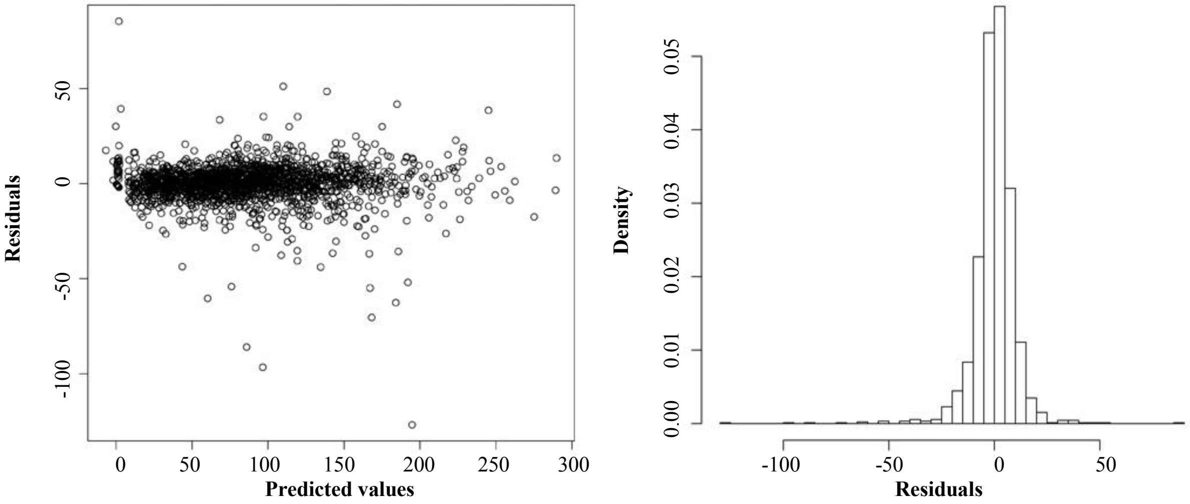

Figure 9.1 EBP residuals against predicted values (left), and histogram of EBP residuals (right).

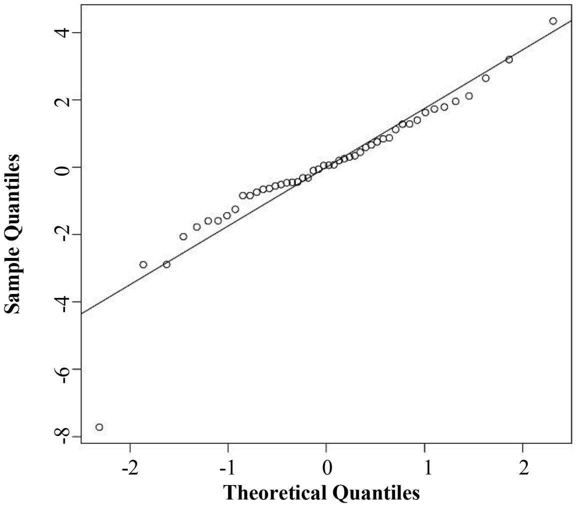

Figure 9.2 Normal Q-Q plot of predicted province effects $\hat{u}_i$.

We proceed now to compare the obtained estimates. Figure 9.3 left shows EBPs of the total sales of the considered tobacco product for each province against direct estimates. Province sample sizes are used as point labels. This plot indicates a great similarity of the two types of estimates except for the two provinces with the largest sample sizes, where the EBPs are slightly larger than direct estimates, which could be due to cut-off sampling bias of the direct estimator. Figure 9.3 right displays EBPs against GREG estimates. The great similarity of GREG and EBP estimates shown by this plot supports the fact that direct estimators might be actually understating the total sales in this application.

Finally, we compare the three types of estimates of the total sales for each province in Figure 9.4 left, showing the point estimates for each province (x-axis), with provinces sorted from smaller to larger sample sizes, and with sample sizes indicated in the x-axis labels. The conclusions are the same as before; that is, the three types of estimates take very similar values for all provinces except for a couple of provinces with the larger sample sizes, where the basic direct estimator takes slightly smaller values (possibly understating the total sales). Figure 9.4 (right) shows the estimated coefficients of variation (CV) obtained ignoring the bias due to cut-off sampling. EBP estimators perform uniformly better than the other estimators in terms of estimated CV, keeping the CV values below 10% for practically all provinces, whereas GREG estimator obtains CV values above 10% for the provinces with the smallest sample sizes. We can see some peaks in the estimated CVs for some provinces with not necessarily the smallest sample sizes. These larger CV values are due to the presence of zero purchases and sales of the considered product in many tobacco shops for those particular provinces (that particular product is not acquired every month). Clearly, the direct estimator performs the worst in terms of efficiency.

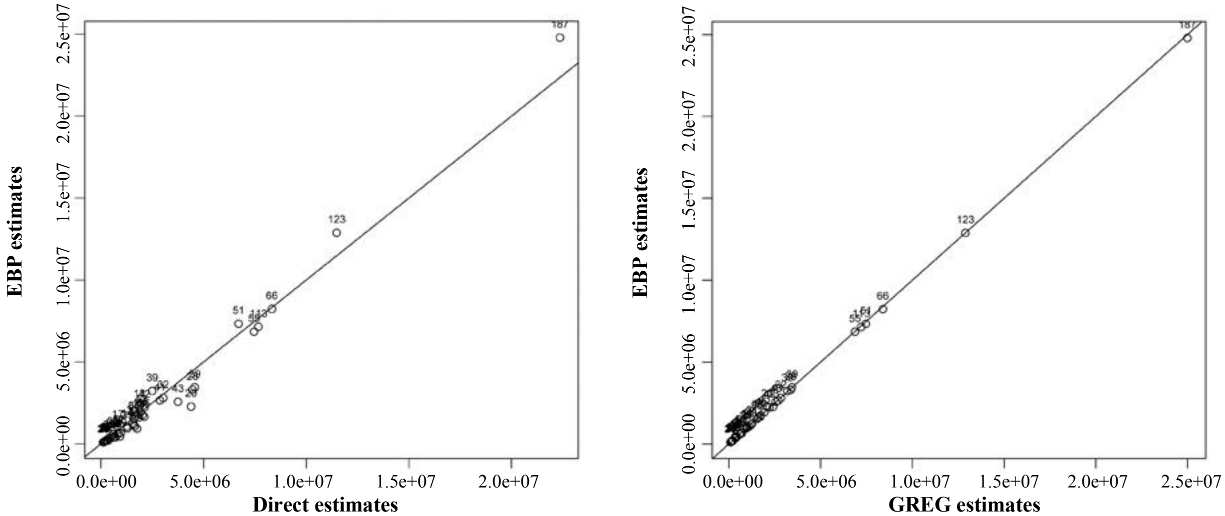

Figure 9.3 EBPs of total sales for each province against direct estimates (left) and against GREG estimates (right).

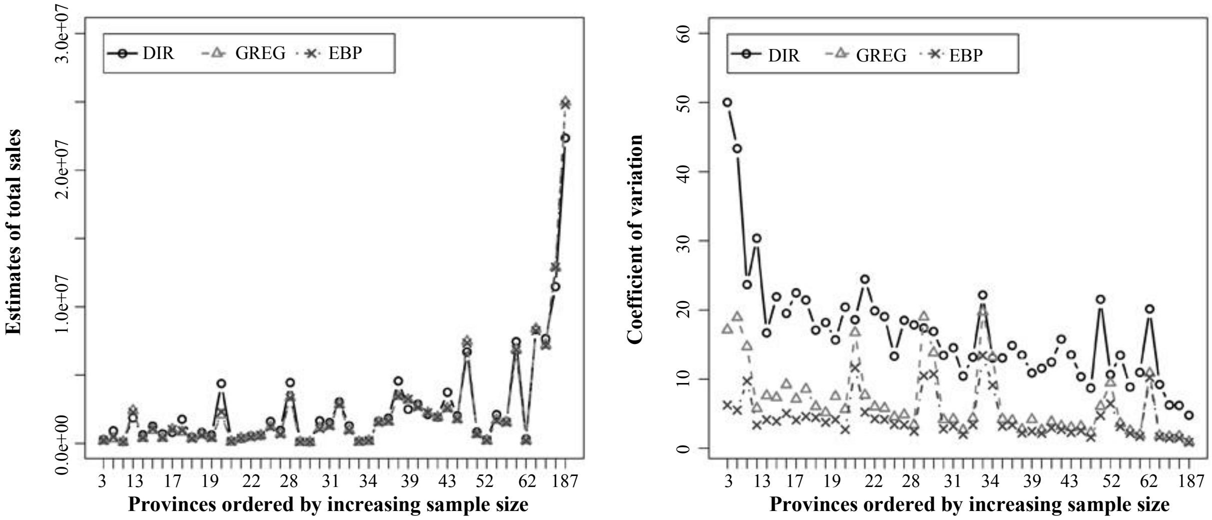

Figure 9.4 Direct, calibration and EBP estimates of total sales for each province (left) and corresponding estimated coefficients of variation (right).

Table A.1 in the Appendix reports direct, LCAL and EBP estimates of province total sales of the product supplemented with their estimated CVs. This table confirms the better performance of EBP in terms of estimated CV under the nested error model, specially for those provinces with small sample sizes. Finally, the direct estimator performs poorly in terms of CV even if the bias due to cut-off sampling is not accounted for.

10 Conclusions

Cut-off sampling is frequently used in business surveys, when drawing a representative sample from the whole population entails a cost that does not really compensate the subsequent gain in accuracy. On the other hand, in some surveys, part of the target population may not be actually available for sampling; that is, there may be population sectors that cannot be represented in the sample. These situations appear more often than desired, providing biased direct estimates as we have seen along this work.

We have studied the theoretical design properties of basic direct, calibration and model-based estimators under cut-off sampling in small areas. Our results show that EBLUP for a linear parameter, similarly as calibration estimators, reduce considerably the bias due to cut-off sampling if the models for the included and excluded individuals are reasonably similar. In terms of MSE, EBLUP performs significantly better than calibration estimators, specially for domains with small sample size.

In our simulation studies and in the application, we compared the proposed methods by assuming that the model is the same for all units in the population (included or excluded). The model assumption could be arguable because there is no way of checking the model for the excluded units. In the case that estimation for the overall domain (and not only for U_{iI}) is required as is the case in this work, one will need to rely on subjective prior information concerning the validity of the assumed model for the excluded units. In any case, estimates can be considered just as indicatives of what could be the true values in the case that the same model holds for all the domain units. In fact, the case of different models for included and excluded units was also analyzed in simulations. In this case, model-based estimators remained to be the most efficient, with not much larger bias than that of calibration estimators.

MSEs of calibration and model-based estimators are obtained under the model. Design MSEs are preferred by National Statistical Institutes because they do not assume that a model is correct and therefore account for model failures. However, finding design-unbiased estimators for the design MSE under cut-off sampling encounters the same problems as finding design-unbiased estimators of the target domain indicators H_i. We plan to use the ideas of Strzalkowska-Kominiak and Molina (2019), based on borrowing strength from the other domains also for estimating the design MSE in a given domain, to find design MSE estimators with reduced cut-off sampling bias.

Finally, we have considered that the domains act as sampling strata and cut-off sampling is applied within each domain. Considering that the strata are different from the domains (typically cutting-across the domains) and applying cut-off sampling within each strata yields random domain sample sizes. Small area estimation is seldom studied under this case in the literature. Nevertheless, putting together the subsamples from the different strata corresponding to the same domain we get a sample from each domain. Inference could then be done conditionally on the observed domain sample sizes Rao (1985), which would reduce to the same problem considered here.

Acknowledgements

The work of M. Guadarrama and I. Molina is supported by the Spanish Ministerio de Economía y Competitividad, grants MTM2015-69638-R (MINECO/FEDER, UE) and MTM2015-72907-EXP.

Appendix

Estimates of total sales by provinces

Table A.1
Basic direct, GREG and EBP estimates of total sales for the selected product and estimated coefficients of variation (%) for each Spanish province (by increasing sample size)

PROVINCE	n_i	$\hat{V}_i^{\text{HA}}$	$\hat{V}_i^{\text{GREG}}$	$\hat{V}_i^{\text{EBP}}$	cv($\hat{V}_i^{\text{HA}}$)	cv($\hat{V}_i^{\text{GREG}}$)	cv($\hat{V}_i^{\text{EBP}}$)
SORIA	3	293,020.0	187,824.9	213,325.0	50.0	17.1	6.2
ZAMORA	7	932,520.0	345,095.8	454,657.0	43.3	18.9	5.5
ALAVA	11	130,083.6	119,918.5	118,835.3	23.7	14.7	9.7
ALMERIA	13	1,870,104.6	2,407,333.1	2,272,051.4	30.4	5.8	3.4
PALENCIA	14	626,340.0	380,367.4	409,775.4	16.7	7.6	4.1
SALAMANCA	14	1,265,580.0	966,094.1	1,068,230.6	21.9	7.3	3.9
AVILA	15	708,696.0	392,474.1	418,917.2	19.5	9.2	5.0
LERIDA	17	817,817.6	1,011,032.3	1,014,770.2	22.5	7.1	4.1
CIUDAD REAL	18	1,764,000.0	841,228.2	939,994.9	21.4	8.6	4.6
GUADALAJARA	18	463,047.8	362,148.3	363,856.9	17.1	6.0	4.5

Table A.1 (continued)
Basic direct, GREG and EBP estimates of total sales for the selected product and estimated coefficients of variation (%) for each Spanish province (by increasing sample size)

PROVINCE	n_i	$\hat{V}_i^{\text{HA}}$	$\hat{V}_i^{\text{GREG}}$	$\hat{V}_i^{\text{EBP}}$	cv($\hat{V}_i^{\text{HA}}$)	cv($\hat{V}_i^{\text{GREG}}$)	cv($\hat{V}_i^{\text{EBP}}$)
RIOJA	18	809,900.0	622,488.3	595,178.6	18.2	5.2	3.7
SEGOVIA	19	610,370.5	386,734.4	402,324.0	15.7	7.5	4.2
CACERES	20	4,391,826.0	2,081,619.7	2,286,462.0	20.4	5.6	2.7
GUIPUZCOA	20	181,634.0	136,700.0	156,311.8	18.6	16.7	11.6
HUESCA	22	377,954.5	372,101.3	371,246.5	24.5	7.7	5.2
TERUEL	22	534,417.3	446,565.7	465,643.3	19.9	6.0	4.3
CUENCA	23	588,464.3	587,005.5	586,347.5	19.0	5.8	4.2
VALLADOLID	24	1,609,875.0	1,210,132.8	1,188,336.1	13.3	4.5	3.4
BURGOS	28	961,645.7	708,510.0	666,698.1	18.5	4.9	3.4
CORDOBA	28	4,457,614.3	3,367,169.5	3,312,801.5	17.9	3.4	2.4
ORENSE	28	148,577.1	88,104.6	108,428.9	17.4	19.0	10.5
LUGO	30	107,213.3	92,938.7	104,233.7	16.9	13.8	10.7
ALBACETE	31	1,654,606.5	1,115,182.2	1,073,719.8	13.4	4.2	2.8
LEON	31	1,528,254.2	1,274,531.6	1,270,341.6	14.5	4.2	3.2
PROVINCE	n_i	$\hat{Y}_i^{\text{DIR}}$	$\hat{Y}_i^{\text{GREG}}$	$\hat{Y}_i^{\text{EBP}}$	cv($\hat{Y}_i^{\text{DIR}}$)	cv($\hat{Y}_i^{\text{GREG}}$)	cv($\hat{Y}_i^{\text{EBP}}$)
HUELVA	32	3,031,328.1	2,838,874.0	2,816,281.3	10.5	2.6	2.0
NAVARRA	33	1,291,343.0	956,737.9	957,660.4	13.2	4.4	3.4
PONTEVEDRA	33	159,229.1	107,198.9	138,367.4	22.2	19.7	13.4
VIZCAYA	34	228,618.8	183,267.3	206,304.6	13.1	13.2	9.1
TOLEDO	35	1,619,939.4	1,529,104.8	1,539,799.3	13.1	4.2	3.2
CADIZ	38	1,851,521.1	1,585,755.9	1,620,844.2	14.9	4.0	3.4
BADAJOZ	39	4,571,743.6	3,439,625.5	3,457,692.5	13.5	2.7	2.2
MALAGA	39	2,499,392.3	3,188,031.1	3,237,081.8	10.9	4.2	2.5
TARRAGONA	41	2,872,882.0	2,690,969.7	2,656,117.8	11.6	2.6	2.2
GRANADA	42	2,123,693.3	2,221,155.1	2,241,916.2	12.5	3.8	2.9
JAEN	43	1,928,229.8	1,940,379.2	1,943,101.0	15.8	3.2	2.7
ZARAGOZA	43	3,750,210.7	2,564,909.0	2,578,011.3	13.5	3.0	2.3
GERONA	45	2,029,222.2	1,748,165.7	1,767,490.3	10.4	3.2	2.5
MURCIA	51	6,700,070.6	7,467,465.0	7,341,434.6	8.7	2.2	1.6
BALEARES	52	849,950.8	650,012.6	694,416.3	21.5	6.1	4.7
CANTABRIA	52	285,632.3	204,947.7	226,163.1	10.7	9.5	6.4
ASTURIAS	55	2,113,034.5	1,702,020.8	1,661,932.8	13.5	3.6	3.1
CASTELLON	55	1,605,604.4	1,526,618.1	1,530,394.2	8.9	2.5	2.2
SEVILLA	55	7,458,078.2	6,878,368.2	6,857,368.8	11.0	2.0	1.7
CORUNA	62	340,200.0	217,028.5	206,041.8	20.2	10.9	10.2
ALICANTE	66	8,324,589.1	8,390,895.3	8,240,996.9	9.2	1.8	1.6
VALENCIA	113	7,671,137.7	7,209,128.2	7,153,290.2	6.3	1.7	1.4
MADRID	123	11,483,342.8	12,892,853.8	12,892,305.0	6.2	1.7	1.5
BARCELONA	187	22,356,500.5	24,990,558.9	24,797,372.9	4.8	1.0	0.9

References

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

Benedetti, R., Bee, M. and Espa, G. (2010). A framework for cut-off sampling in business survey design. *Journal of Official Statistics*, 26, 651-671.

Foster, J., Greer, J. and Thorbecke, E. (1984). A class of decomposable poverty measures. *Econometrica: Journal of the Econometric Society*, 761-766.

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

Hájek, J. (1971). Discussion of an essay on the logical foundations of survey sampling, comment on a paper by D. Basu. In *Foundations of Statistical Inference*, (Eds., V.P. Godambe and D.A. Sprott), Toronto: Holt, Rinehart, Winston.

Haziza, D., Chauvet, G. and Deville, J.-C. (2010). Sampling estimation in presence of cut-off sampling. *Australian & New Zealand Journal of Statistics*, 52, 303-319.

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

INE (2018). Índices de producción industrial (IPI) base 2015. Technical report, Instituto Nacional de Estadística, España.

Molina, I., and Marhuenda, Y. (2015). sae: An R package for small area estimation. *R Journal*, 1, 81-98.

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

Pratesi, M. (2016). *Analysis of Poverty Data by Small Area Estimation*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Rao, J.N.K. (1985). Conditional inference in survey sampling. *Survey Methodology*, 11, 1, 15-31. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/1985001/article/14364-eng.pdf.

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

R Development Core Team (2016). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna.

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

Strzalkowska-Kominiak, E., and Molina, I. (2019). Estimation of proportions in small areas: Application to the labour force using the Swiss Census Structural Survey. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*.

Tillé, Y., and Matei, A. (2016). *Sampling: Survey Sampling*. R package version 2.8.

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

Survey Methodology, June 2020
Vol. 46, No. 1, pp. 77-91
Statistics Canada, Catalogue No. 12-001-X

Model-assisted sample design is minimax for model-based prediction

Robert Graham Clark[1]

Abstract

Probability sampling designs are sometimes used in conjunction with model-based predictors of finite population quantities. These designs should minimize the anticipated variance (AV), which is the variance over both the superpopulation and sampling processes, of the predictor of interest. The AV-optimal design is well known for model-assisted estimators which attain the Godambe-Joshi lower bound for the AV of design-unbiased estimators. However, no optimal probability designs have been found for model-based prediction, except under conditions such that the model-based and model-assisted estimators coincide; these cases can be limiting. This paper shows that the Godambe-Joshi lower bound is an *upper* bound for the AV of the best linear unbiased estimator of a population total, where the upper bound is over the space of all covariate sets. Therefore model-assisted optimal designs are a sensible choice for model-based prediction when there is uncertainty about the form of the final model, as there often would be prior to conducting the survey. Simulations confirm the result over a range of scenarios, including when the relationship between the target and auxiliary variables is nonlinear and modeled using splines. The AV is lowest relative to the bound when an important design variable is not associated with the target variable.

Key Words: Anticipated variance; Model-based inference; Probability sampling; Sample surveys.

1 Introduction

Model-based inference about finite population totals relies on an assumed model and usually does not make reference to the sampling plan. Probability sampling, where every unit i has a known probability of selection $\pi_i > 0,$ is not strictly necessary, but is often used anyway, because it "eliminates conscious and unconscious bias" (Valliant, Dever and Kreuter, 2013, page 310) and ensures the non-informativeness of sampling which is required for most model-based procedures (Chambers and Clark, 2012, page 12). Särndal, Swensson and Wretman (1992, page 534) note that "proponents of model-based inference advocate randomized selection of the sample as a safeguard against selection bias, but the randomization probabilities play no role in the inference". See also Lohr (2010, page 263), Chambers and Clark (2012, page 92) and Scott, Brewer and Ho (1978) who suggest probability sample designs for model-based predictions. For a review of the model-based approach, see also Valliant, Dorman and Royall (2000).

Model-assisted inference (e.g., Särndal et al., 1992) is an alternative approach, where estimators are design-unbiased (at least asymptotically), that is, unbiased over repeated probability sampling from any fixed population. Subject to this constraint, they minimize the anticipated variance (AV), which is the variance over both repeated realisations of the population from a model and repeated probability sampling. In models with independent errors, the lowest possible AV amongst such estimators (for any given probability sample design) is the Godambe-Joshi lower bound (GJLB) (Godambe and Joshi, 1965). The lower bound is asymptotically achieved for linear models by the well known generalized regression estimator.

[1] Robert Graham Clark, Research School of Finance, Actuarial Studies and Statistics, Australian National University, Canberra, Australia. E-mail: robert.clark@anu.edu.au.

Model-assisted designs are probability sample designs which are intended to minimize the AV of the generalized regression estimator (or, equivalently, to minimize the GJLB). These AV-optimal sample designs have been derived for model-assisted inference. In particular, the sample design which minimizes the GJLB for fixed expected sample size for models with independence has probability proportional to the square root of the model error variance for each unit (σ_i) (e.g., Särndal et al., 1992); this will be called a PPσ design. The PPCσ design is a generalization allowing for unequal unit costs (Steel and Clark, 2014). There are no analogous results on optimal probability sampling for model-based prediction, except under strong conditions, a gap that this paper partially fills. Isaki and Fuller (1982) suggested the design-estimation strategy of using the PPσ design for model-based prediction, showing that this design is optimal when selection probabilities and their squares are in the column space of the matrix of covariates. This condition comes at a price, as will be seen in the simulation in this paper.

Optimal non-probability samples have been derived for model-based best linear unbiased predictors (BLUPs) under linear models. These tend to be somewhat extreme designs, where the units with the largest, or the largest and smallest, values of auxiliary variables are chosen (e.g., Royall, 1970). Robust model-based balanced designs have been developed, where one or more sample moments of auxiliary variables are equal to the corresponding population moments (Royall and Herson, 1973), while "over-balanced designs" meet a different constraint on the sample moments (Scott et al., 1978). Another balanced design was proposed by Kott (1986). These designs are robust to families of polynomial alternatives to a working linear model. They are not probability designs, although probability designs have been proposed in order to approximately meet balancing constraints (Valliant et al., 2000, Section 3.4). Exactly balanced probability designs have also been proposed (Tillé, 2006). The choice of balancing or over-balancing strategy depends on which set of polynomial alternatives is postulated. In another non-probability approach, Welsh and Wiens (2013) find the sample which minimizes the maximum model-based variance in a neighbourhood of a working model.

This article derives an asymptotic upper bound for the AV of the BLUP under probability sampling. The AV is the most relevant quantity for probability sample design even in the model-based framework, because averaging over all possible samples is appropriate in advance of sample selection. The bound is applicable to any probability sample design, and is over the space of possible covariate sets. This is useful for sample design in practice, because the precise model to be used is not decided until after data have been collected. For example, some design variables might not be included in the model if the sample data suggests that they have little relevance for the variable whose total is being estimated, but this would not be clear prior to surveying. Or splines might be used, with the number and placement of knots guided by the sample data. It turns out that the upper bound is the GJLB. This implies that model-assisted designs, such as PPσ and PPCσ, are minimax strategies for model-based estimation. The upper bound is an equality when the model has a particular property, which is satisfied when the model is sufficiently rich and includes all design variables.

Other researchers have considered the relationship between the BLUP and the model-assisted generalized regression estimator, including conditions under which these two estimators are identical (e.g., Isaki and Fuller, 1982; Tam, 1988) and modifications to the BLUP so that it is equivalent to a generalized regression estimator at the expense of its optimality under the model (e.g., Brewer, Hanif and Tam, 1988; Brewer, 1999; Nedyalkova and Tillé, 2008). The results here are new because:

- Existing results do not cater for situations where both: the surveyor wants to use the BLUP because it is model-optimal, and the BLUP and the generalized regression estimators are not equal. The case where the two estimators are equal is shown here to be, in a particular sense, the worst case for the BLUP.
- An expression for the AV of the BLUP is derived and shown explicitly to be less than or equal to the upper bound. The result seems intuitively reasonable, given that the GJLB is attained by the design-consistent generalized regression estimator, whereas the BLUP is not subject to the constraint of design-based consistency. However, it is not at all obvious from the expression for the BLUP's AV that the upper bound applies, so it is useful to have an explicit result.
- The interpretation is made that the upper bound is over the space of possible choices for the model covariates $\mathbf{x}$. Thus, the upper bound is relevant when the sample designer is unsure what model will ultimately be adopted once data have been collected.

Section 2 contains the key theoretical results. Section 3 confirms and illustrates the main result in a simulation study with a variable of interest Y and two auxiliary variables: x_1 (continuous) and x_2 (binary). The expected value of Y conditional on these variables is defined by a linear and a sinusoidal term in x_1. It does not depend on x_2. The probabilities of selection are a function of both x_1 and x_2. BLUPs are calculated based on the model with lowest Bayesian Information Criterion (BIC) from a set including the simple linear model in x_1 and splines in x_1 of various degrees, both with and without x_2. The ratio of the simulation prediction mean squared error (MSE) of the BLUP to the GJLB is either less than or equal to 1 or just above 1 across a range of scenarios. Section 4 is a discussion.

Much of the literature comparing model-assisted and model-based estimators and inference has focussed on bias due to mis-specified models when either (a) the mean function is incorrect, or (b) some design variables are inappropriately excluded. See for example Hansen, Madow and Tepping (1983) and the reworking of their simulation study in Valliant et al. (2000, Section 3.4). The simulation in Section 3 considers (a) and (b) to some extent, but this isn't the main focus of the paper. The aim here is to see whether a committed model-based statistician can use the GJLB as an upper bound for the AV for sample design purposes, rather than to adjudicate between model-based and design-based inference. It is assumed that a sufficiently good model can be identified using the sample data; this process would be aided by a design which minimizes the maximal AV over the space of all linear models. A PPσ or PPCσ design is recommended when there is considerable uncertainty over the form of the final model.

2 Upper bound for the AV of the BLUP

Let $U = \{1, \ldots, N\}$ denote the finite population. For unit $i \in U$, the variable of interest is y_i and the p-vector of auxiliary variables is $\mathbf{x}_i$. The sample (of size n) is s and the non-sample set is $r = U - s$. Auxiliary variables are observed for all $i \in U$ while y_i is observed for $i \in s$. The aim is to predict $t_y = \sum_{i \in U} y_i$. The probabilities of selection are $\pi_i = P[i \in s \,|\, \mathbf{x}_1, \ldots, \mathbf{x}_N, y_1, \ldots, y_N] = P[i \in s \,|\, \mathbf{x}_1, \ldots, \mathbf{x}_N] > 0$; they are assumed to be a function of the population values of $\mathbf{x}_i$. Let $\mathbf{t}_x = \sum_{i \in U} \mathbf{x}_i$ and $\mathbf{t}_{xr} = \sum_{i \in r} \mathbf{x}_i$.

The n by p matrix of sample values of $\mathbf{x}$, which has rows $\mathbf{x}_i^T$, is denoted X_s. The $N - n$ by p matrix of non-sample values of $\mathbf{x}$ is X_r. The vector of sample values of y is $\mathbf{y}_s$.

The following linear model M is assumed:

$$E_M[y_i] = \boldsymbol{\beta}^T \mathbf{x}_i \tag{2.1}$$

$$\operatorname{var}_M[y_i] = \sigma_i^2 = \sigma^2 v_i \tag{2.2}$$

$$\operatorname{cov}_M\left[y_i, y_j\right] = 0 \tag{2.3}$$

for $i, j \in U$ with $i \neq j$. The subscripts M in E_M, var_M and cov_M indicate distributions over repeated realisations of the population values from the model. It is generally assumed that v_i are known, i.e., the error variances are known up to a constant of proportionality. For example, in business surveys, v_i might be a measure of business size, or the square root thereof. The unknown parameters are $\boldsymbol{\beta}$ and σ^2. The values of $\mathbf{x}_i$ are considered to be fixed.

The best linear unbiased predictor (BLUP) (denoted $\hat{t}_y$) for a generalization of model M is stated in Chapter 2 of Valliant et al. (2000). Its model-based prediction variance is

$$\operatorname{var}_M\left(\hat{t}_y - t_y\right) = \mathbf{t}_{xr}^T \left(\sum_{i \in s} \sigma_i^{-2} \mathbf{x}_i \mathbf{x}_i^T\right)^{-1} \mathbf{t}_{xr} + \sum_{i \in r} \sigma_i^2. \tag{2.4}$$

(This can be obtained as a special case of Result 2.2.2 on page 29 of Valliant et al., 2000.)

The anticipated variance is defined as $\operatorname{AV}\left(\hat{t}_y - t_y\right) = E_M E_p \left(\hat{t}_y - t_y\right)^2$ (Isaki and Fuller, 1982). As $\hat{t}_y$ is model-unbiased, its AV is equal to

$$\operatorname{AV} = E_p \operatorname{var}_M \left(\hat{t}_y - t_y\right).$$

Theorem 1 will derive an approximation for this AV. The asymptotic framework is based on the design-based asymptotics of Isaki and Fuller (1982). It is assumed that there is a countably infinite population $i = 1, 2, \ldots$. A sequence of finite populations U_t is defined by $U_t = \{1, \ldots, N_t\}$ where $N_1 < N_2 < \ldots$. For each t, a sample s_t of size n_t is selected from U_t by arbitrary probability design with probabilities of selection $\pi_{i(t)} = P\left[i \in s_t\right]$. Following (2.7) of Isaki and Fuller (1982), it is assumed that

$$0 < \lambda_1 < \pi_{i(t)} < \lambda_2 \tag{2.5}$$

for some constants λ_1 and λ_2. Isaki and Fuller (1982) note that AVs of estimators of total are typically $O(n_t)$ (which is equivalent to $O(N_t)$) and the totals themselves are also $O(n_t)$. Population means will be denoted as $\bar{\mathbf{X}}_t = N_t^{-1}\sum_{U_t}\mathbf{x}_i$ and the inverse probability weighted estimator of $\bar{\mathbf{X}}$ is $\hat{\bar{\mathbf{X}}}_\pi = N_t^{-1}\sum_{s_t}\pi_{i(t)}^{-1}\mathbf{x}_i$ (and similarly for Y and other variables).

Two new variables are defined for each unit i by $\mathbf{u}_{i(t)} = \pi_{i(t)}\mathbf{x}_i$ (a p-vector) and $\mathbf{v}_{i(t)} = \pi_{i(t)}\sigma_i^{-2}\mathbf{x}_i\mathbf{x}_i^T$ (a p by p matrix). Their population means are $\bar{\mathbf{U}}_t$ and $\bar{\mathbf{V}}_t$ with inverse probability estimators $\hat{\bar{\mathbf{U}}}_{t\pi}$ and $\hat{\bar{\mathbf{V}}}_{t\pi}$.

Theorem 1. *It is assumed that*

$$\lim_{t\to\infty} E_p\left\{\hat{\bar{\mathbf{U}}}_{t\pi}^T\hat{\bar{\mathbf{V}}}_{t\pi}^{-1}\hat{\bar{\mathbf{U}}}_{t\pi} - \bar{\mathbf{U}}_t^T\bar{\mathbf{V}}_t^{-1}\bar{\mathbf{U}}_t\right\} = 0. \tag{2.6}$$

Then

$$E_p var_M\left(\hat{t}_y - t_y\right) = AV + o(n_t). \tag{2.7}$$

where

$$AV = \sum_U (1-\pi_i)\mathbf{x}_i^T\left(\sum_U \pi_i\sigma_i^{-2}\mathbf{x}_i\mathbf{x}_i^T\right)^{-1}\sum_U (1-\pi_i)\mathbf{x}_i + \sum_U (1-\pi_i)\sigma_i^2. \tag{2.8}$$

Notes on Theorem 1

- Assumption (2.6) is reminiscent of Result (3.24) of Isaki and Fuller (1982), but there is an important difference. In Isaki and Fuller (1982), unit variables depend only on i, but here $\mathbf{u}_{i(t)}$ and $\mathbf{v}_{i(t)}$ depend on both i and t as they both have a factor $\pi_{i(t)}$. However, $\pi_{i(t)}$ are bounded by (2.5), so the condition is plausible; it would not be if $\pi_{i(t)}$ could be arbitrarily close to zero.
- It is clear that assumption (2.6) is satisfied if $\hat{\bar{\mathbf{U}}}_{t\pi}$ and $\hat{\bar{\mathbf{V}}}_{t\pi}$ are consistent in design probability for $\bar{\mathbf{U}}_t$ and $\bar{\mathbf{V}}_t$, and $\hat{\bar{\mathbf{V}}}_{t\pi}$ is invertible in a neighbourhood of $\bar{\mathbf{V}}_t$. As noted by Isaki and Fuller (1982) in a comment on their condition (3.12), a invertibility requirement of this sort seems reasonable "for any discussion of regression estimation".

An upper bound for the asymptotic AV over all possible choices of the auxiliary vector $\mathbf{x}_i$ will now be derived. This allows for uncertainty about which auxiliary variables will ultimately be included in the model, since this decision is typically only made after data is collected. For example, the full set of variables used in the design might or might not end up being in the model, or spline functions of covariates might be included with knots based on the sample data. Theorem 2 states the upper bound.

Theorem 2. *Let AV be the asymptotic AV defined by (2.8). If $\sum_U \pi_i\sigma_i^{-2}\mathbf{x}_i\mathbf{x}_i^T$ is invertible and $\pi_i > 0$ for all $i \in U$, then*

$$AV \le \sum_U \left(\pi_i^{-1} - 1\right)\sigma_i^2 \tag{2.9}$$

with strict equality if and only if there exists a p *-vector* $\boldsymbol{\lambda}$ *such that*

$$\left(\pi_i^{-1} - 1\right) \sigma_i^2 = \boldsymbol{\lambda}^T \mathbf{x}_i \qquad (2.10)$$

for all $i \in U$.

The right hand side of (2.9) is the well known Godambe-Joshi lower bound (Godambe and Joshi, 1965) for the AV of design-unbiased estimators. Here it is an upper bound over model space for model-based BLUPs.

Suppose the total cost of running the survey is $\sum_s C_i$ plus fixed costs, where C_i is the cost associated with surveying unit i. Then the expected cost is $C_E = \sum_U C_i \pi_i$. The sample design which minimizes the upper bound in (2.9) subject to fixed cost is the PPCσ design which has

$$\pi_i \propto \sigma_i \big/ \sqrt{C_i} \qquad (2.11)$$

(Steel and Clark, 2014 who generalize Särndal et al., 1992, page 452) to allow for unequal costs. Theorem 2 means that (2.11) is a minimax design when there is uncertainty about the form of the model. Note that only the first order inclusion probabilities affect the AV and the bound, but these do not fully specify the design. Samples can be selected using these inclusion probabilities in a variety of ways (Tillé, 2006), including balanced probability sampling (Nedyalkova and Tillé, 2012) which improves the robustness to model mis-specification.

The condition for equality, (2.10), is equivalent to a well known condition for the BLUP to be equal to the generalized regression estimator (formula 3 of Tam, 1988). Tam (1988) argued for the use of sample designs such that (2.10) is satisfied, such as PPσ (provided that the model includes an intercept). Nedyalkova and Tillé (2008), building on a result from Royall (1992), showed that PPσ is model-based-optimal under equal costs when both v_i and $\sqrt{v_i}$ are linear functions of $\mathbf{x}_i$, a condition called *explainable variances*. Brewer et al. (1988) noted that (2.10) can also be satisfied if the estimation model includes an instrumental variable, which is a suitable function of the selection probabilities. However, there are many circumstances under which (2.10) is not satisfied, because some auxiliary variables are omitted from the final model, because multiple variables of interest have different variance structures (ruling out PPσ), because there are unequal costs, or because instrumental variables are eschewed due to the loss of efficiency they entail. Theorem 2 shows that the GJLB is an upper bound under these circumstances which are not covered by the results of these authors. The PPCσ design in (2.11) is a minimax design in this more general setting.

3 Simulation study

A simulation study was conducted to compare the AV of the BLUP and its upper bound in situations where Y has a nonlinear relationship with a continuous auxiliary variable x_1. A second auxiliary variable, x_2, is binary and independent of x_1 and Y. Probabilities of selection depend on x_1 and x_2 in various ways. For each scenario, 5,000 populations and samples were generated, with a population size of 6,000 and sample sizes of 500 and 1,500, and with a population size of 100,000 and a sample size of 25,000. All code is available at www.github.com/rgcstats/AVLB.

3.1 Simulation of populations

The population values of x_1 were the $\{j/(101):\ j = 1, \ldots, 100\}$ quantiles of a lognormal distribution with mean $-1/32$ and standard deviation 0.25, with equal frequencies of each of these 100 values. This means that x_1 are positive and right-skewed with a mean of 1 and a range of 0.54 to 1.74. The values of x_1 were non-stochastic and discretized in order to speed up computation, to simplify the generation of smooth models (see below), and to facilitate comparison of AVs to the GJLB by making the GJLB constant across simulations. A second binary variable x_2 took on the values 0 and 1 with equal frequency within each value of x_1, so that the two covariates were orthogonal.

Conditional on x_1 and x_2, the population values of Y were generated independently as

$$E_M Y_i = \mu(x_{1i}) + \varepsilon_i \tag{3.1}$$

where

$$\varepsilon_i \sim N(0,\ 0.25 x_{1i}). \tag{3.2}$$

The mean function $\mu(.)$ was a smooth but nonlinear function,

$$\mu(x_1) = 4x_1 + \sin(x_1 2\pi/h), \tag{3.3}$$

consisting of a linear term and a sinusoidal term with period h. When h is large, $\mu()$ is close to linear over the range of x_1 in the population, while for h small there are frequent cycles in the function. Figure 3.1 shows the mean function for the periods used in the simulation (0.5, 1, 2, 5).

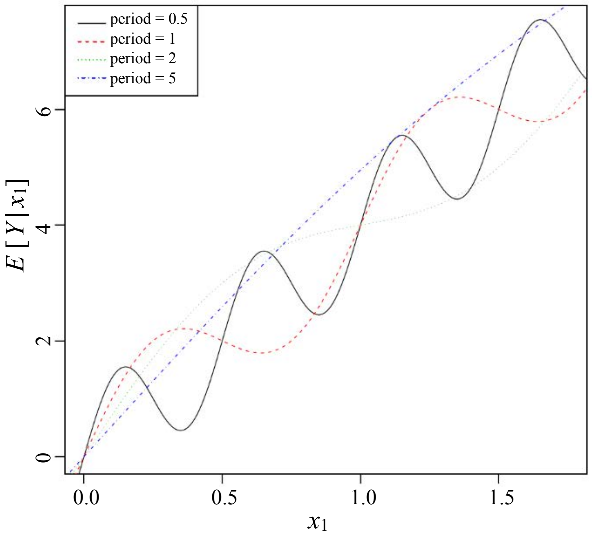

Figure 3.1 $\mu(x_1) = E[Y \mid x_1]$ **for the periods used in the simulation study.**

3.2 Simulated sampling

The probabilities of selection, π_i, were set to

$$\pi_i \propto x_{1i}^b (1 + cx_{2i}) \tag{3.4}$$

where b was 0.5, 1 or 2, reflecting light, medium or high dependence on x_{1i}. The values of c were 0, 0.5 or 1.5, reflecting no, medium or high dependence on x_{2i}. (Other values of c were also used for the purposes of Figure 3.2 only.)

The second auxiliary variable, x_2, is unrelated to Y but may affect the selection probabilities. One might expect the BLUP to do better compared to the GJLB when probabilities of selection depend on x_2, since the BLUP may be based on a model omitting x_2, potentially leading to lower variance. Of course, there is also the possibility of the working model omitting x_1, leading to the BLUP being biased, however this never occurred in any of the simulations. To explore robustness to incorrect omission of x_1, it would be of interest to consider relationships weaker than those shown in Figure 3.1, but this was beyond the scope of the paper.

Inclusion probabilities are forced to obey the proportionality in (3.4) but are truncated above at 1 and below at 1/40 and scaled such that they add to the required sample size after truncation. Samples are selected by unequal probability systematic sampling with random ordering using the `sampling` package in R (Tillé and Matei, 2016).

3.3 Estimation of the population total of *Y*

A linear model in x_1 and spline models in x_1 with between 1 and 10 interior knots are fitted to each sample. (See for example Breidt, Claeskens and Opsomer, 2005 for the use of splines in model-assisted survey estimation.) Another 11 models are defined by also including x_2 as an additive covariate. The model with the lowest BIC is then used to calculate a BLUP of t_y. This model selection step would be expected to increase the variability of this predictor. The BLUP based on the simple linear model in x_1 is also calculated. The process is repeated with working models including the correct variance specification $\text{var}_M(y_i) \propto x_{1i}$ and a mis-specification $\text{var}_M(y_i) \propto x_{1i}^2$.

3.4 Simulation results

Tables 3.1-3.4 show the ratios of the prediction MSEs of various BLUP estimators to the GJLB (the right hand side of equation 2.9) for various sample designs and choices of $\mu(x_1)$. The prediction MSEs are the means over all simulations of $(\hat{t}_y - t_y)^2$ so they are with respect to both model and design, as are the results in Theorems 1 and 2.

Table 3.1a evaluates the BLUP corresponding to the lowest-BIC model for a sample size of 500 with correctly specified variances. Nine sample designs are shown corresponding to three choices for b (low/medium/high dependency of the selection probabilities on x_1) and c (no/some/high dependency of selection probabilities on x_2). The period h of the sinusoidal component of $\mu(x_1)$ is also shown (see equation 3.3). The table shows that:

- The ratio is always either less than or equal to 1 or slightly above 1, consistent with Theorem 2. Its range is 0.815 to 1.086.
- The ratio decreases slightly as h increases. So, when the true $E_M(y \mid x_1)$ is close to linear, the model-based BLUP has lower MSE relative to the GJLB, while for more nonlinear models the ratio is closer to 1.
- The ratio depends on b to a degree, although the pattern depends on the other parameters.
- The ratio decreases dramatically as c increases, with reductions of up to 20% from $c = 0.$ This shows that the BLUP does much better relative to the GJLB when there is a covariate x_2 which is relevant in the design but not relevant to Y so that it can be omitted from the estimation model.

Table 3.1b shows results for a much larger sample size of 25,000 from a population of 100,000. These results were included to see whether the ratios are less than or equal to 1 for large n as predicted by the theory in Section 2. A larger number of simulations were also used for this panel (15,000 rather than 5,000). Results are only shown for $c = 0$ since these were the designs with the highest ratios in Table 3.1a. The ratios in 3.1b range from 0.984 to 1.011. The values slightly above 1 may reflect that the working spline model does not perfectly capture the sinusoidal functions used to generate the data.

Table 3.2 shows the same scenarios as Table 3.1a except that the BLUP is based on a mis-specified variance model with $\sigma_i^2 \propto x_i^2$ (the generating model has $\sigma_i^2 \propto x_i$). The ratios of the MSE of the lowest-BIC-model BLUP to the GJLB are on the whole slightly higher than Table 3.1 (generally by less than one percentage point). The ratio is still almost always less than or equal to 1, with a maximum value of 1.089.

Table 3.3 is similar to Table 3.1a except that the sample size is 1,500 rather than 500. The ratios are almost always lower than in Table 3.1a. The maximum ratio is 1.030.

Table 3.4 shows results for the BLUP based on the simple linear model containing only x_1 with mis-specified variance. The sample size is 500. When the period is 5, so that the true model is virtually linear in x_1, this BLUP does very well. The ratios are then always below 1.1 and can be as low as 0.790. When the period is 2, there is visible curvature in $E(y \mid x_1)$ (as shown in Figure 3.1), but the simple BLUP still does well, with all ratios less than 1.2. However, for periods 0.5 and 1, the ratios are well above 1, with a maximum value of 3.4. This shows the substantial bias of the BLUP when the model is badly mis-specified.

The extent to which the selection probabilities depend on x_2 is the major factor determining the ratio of the MSE to the GJLB, as shown by Tables 3.1-3.3. Figure 3.2 shows this phenomenon in more detail for correctly specified variance. The sample size is 1,500 with PPσ sampling so that $\pi_i \propto x_{1i}^{0.5}$. Values of c (0, 0.25, ..., 3) are on the x-axis and results are shown for different periods h. The figure shows that the ratio is slightly above 1 for $c = 0$ and decreases smoothly with c, to about 0.6 when $c = 3$. Higher periods h (reflecting a smoother relationship between Y and x_1) are also associated with lower ratios, but the differences are so small as to be almost indiscernible in Figure 3.2.

Table 3.1
Ratios of MSE of BLUP based on lowest BIC spline model to Godambe-Joshi Lower Bound for sample sizes of 500 and 25,000 with variance correctly specified. Probabilities of selection are proportional to $x_1^b\left(1 + cx_2\right)$. The period h controls the smoothness of $E\left[Y \mid x_1\right]$

		(a) sample size of 500			
sample design		**period (h)**			
b	***c***	**0.5**	**1**	**2**	**5**
0.5	0	1.086	1.064	1.052	1.057
0.5	0.5	0.990	0.963	0.951	0.956
0.5	1.5	0.840	0.827	0.808	0.815
1	0	1.033	1.015	0.997	1.006
1	0.5	1.006	0.992	0.973	0.985
1	1.5	0.877	0.859	0.854	0.858
2	0	1.080	1.063	1.035	1.081
2	0.5	1.046	1.021	0.996	1.039
2	1.5	0.870	0.853	0.839	0.856
		(b) sample size of 25,000			
0.5	0	1.011	1.011	1.011	1.010
1	0	1.007	1.006	1.006	1.006
2	0	0.985	0.985	0.984	0.984

Table 3.2
Ratios of MSE of BLUP based on lowest BIC model to Godambe-Joshi Lower Bound for sample size of 500 with variance mis-specified. Probabilities of selection are proportional to $x_1^b\left(1 + cx_2\right)$. The period h controls the smoothness of $E\left[Y \mid x_1\right]$

sample design		**period (h)**			
b	***c***	**0.5**	**1**	**2**	**5**
0.5	0	1.089	1.068	1.055	1.061
0.5	0.5	0.996	0.967	0.959	0.962
0.5	1.5	0.852	0.830	0.819	0.825
1	0	1.033	1.012	0.998	1.001
1	0.5	1.006	0.992	0.978	0.988
1	1.5	0.886	0.863	0.854	0.862
2	0	1.078	1.061	1.035	1.047
2	0.5	1.048	1.017	0.997	1.010
2	1.5	0.878	0.857	0.842	0.856

Table 3.3
Ratios of MSE of BLUP based on lowest BIC model to Godambe-Joshi Lower Bound for sample size of 1,500 with variance correctly specified. Probabilities of selection are proportional to $x_1^b\left(1 + cx_2\right)$. The period h controls the smoothness of $E\left[Y \mid x_1\right]$

sample design		**period (h)**			
b	***c***	**0.5**	**1**	**2**	**5**
0.5	0	1.030	1.023	1.019	1.022
0.5	0.5	0.928	0.920	0.918	0.922
0.5	1.5	0.762	0.754	0.750	0.749
1	0	0.940	0.936	0.930	0.932
1	0.5	0.979	0.972	0.966	0.968
1	1.5	0.821	0.817	0.810	0.809
2	0	1.028	1.008	0.995	1.020
2	0.5	0.962	0.948	0.941	0.969
2	1.5	0.798	0.798	0.786	0.804

Table 3.4
Ratios of MSE of BLUP based on simple linear model to Godambe-Joshi Lower Bound for sample size of 500 with variance mis-specified. Probabilities of selection are proportional to $x_1^b(1 + cx_2)$. The period h controls the smoothness of $E[Y \mid x_1]$

sample design		period (h)			
b	c	0.5	1	2	5
0.5	0	3.083	2.037	1.109	1.055
0.5	0.5	2.918	1.860	1.002	0.958
0.5	1.5	2.452	1.533	0.840	0.790
1	0	3.086	1.812	1.052	1.007
1	0.5	3.006	1.792	1.016	0.979
1	1.5	2.537	1.500	0.880	0.838
2	0	3.423	2.900	1.174	1.080
2	0.5	3.243	2.693	1.111	1.025
2	1.5	2.689	2.291	0.926	0.829

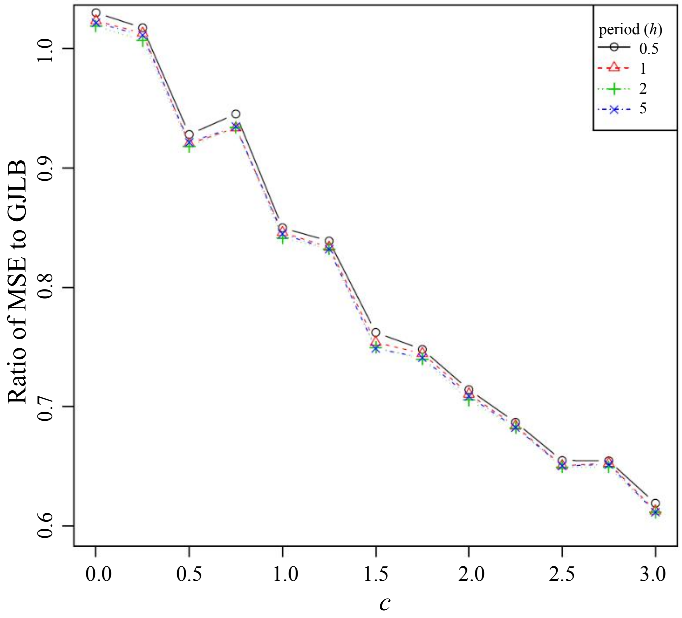

Figure 3.2 Ratios of MSE of BLUP based on lowest BIC model to Godambe-Joshi Lower Bound for sample size of 1,500 with variance correctly specified, vs c, where probabilities of selection are proportional to $x_1^b(1 + cx_2)$ with $b = 1$. The period (h) controls the smoothness of $E[Y \mid x_1]$.

4 Discussion

The Godambe-Joshi lower bound is shown here to be an *upper* bound for the AVs of BLUPs based on a correct model. Simulation MSEs of BLUPs based on an adaptively chosen spline or linear model are consistently less than the GJLB or just above it, even when variances are mis-specified. The MSEs are well below the bound if an important design variable does not figure in the model.

The upper bound result relies on the BLUP being model-unbiased. BLUPs based on a badly mis-specified model had MSEs well above the bound in the simulation study. Choosing a working model with minimum BIC out of a class including spline models avoided this problem.

Once data are available, model-based inference conditions on the sample selected. Probability sampling nevertheless has many advantages even though it is not the basis of inference (e.g., Särndal et al., 1992; Valliant et al., 2013). At the design stage, the AV is then the most relevant objective, because it averages over all the possible samples which may then be selected. The upper bound derived here is relevant at the design stage because there would usually be considerable uncertainty about the form of the model which will ultimately be adopted (there may be exceptions when there are historical or related data to support specification of a model or where one is willing to trust that the true model lies in a class of polynomial or other specific alternative models).

A sensible strategy in practice would be:

i. Set π_i so as to give low values for the upper bound $\sum_{i \in U} \left(\pi_i^{-1} - 1\right) v_i$ where the model variances are proportional to v_i (or a weighted combination of the upper bounds for multiple variables of interest) while also respecting cost and practical considerations. If there is a single variable of interest, and unit costs are proportional to C_i, then $\pi_i \propto \sqrt{v_i / C_i}$ is recommended as it is a minimax strategy.
ii. Once the sample is selected and data are available, choose a regression model based on this data.
iii. Estimate population totals using the BLUPs under the selected model.
iv. This may or may not result in condition (2.10) being satisfied, depending on the costs C_i and the auxiliary variables selected in the final model.

All optimal sample design results, whether model-based, design-based or model-assisted, rely on knowledge of the relative unit or stratum residual variances. This appears to be unavoidable. This paper helps when the form of the mean model is not known in advance by giving an upper bound over the space of models for the mean. There does not appear to be a correspondingly useful bound over possible variance models, so the form of the variance model must be guessed or assumed.

Acknowledgements

I would like to thank Stephen Haslett for his encouragement of methodological research amongst competing priorities. Thorough reviews by two referees and an associate editor substantially improved the paper. This research was undertaken with the assistance of resources from the National Computational Infrastructure (NCI Australia), an NCRIS enabled capability supported by the Australian Government.

Appendix

Proof of Theorem 1

From (2.4),

$$E_p \mathrm{var}_M \left(\hat{t}_y - t_y\right) = E_p \left\{ \mathbf{t}_{xr}^T \left(\sum_{i \in s} \sigma_i^{-2} \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \mathbf{t}_{xr} \right\} + \sum_{i \in U} (1 - \pi_i)\, \sigma_i^2. \tag{A.1}$$

Making use of the definitions of $\hat{\bar{\mathbf{U}}}_\pi$ and $\hat{\bar{\mathbf{V}}}_\pi$, and assumption (2.6), the first term of (A.1) becomes

$$\begin{aligned} E_p\left\{\mathbf{t}_{xr}^T\left(\sum_{i\in s}\sigma_i^{-2}\mathbf{x}_i\mathbf{x}_i^T\right)^{-1}\mathbf{t}_{xr}\right\} &= E_p\left\{\left(N_t\bar{\mathbf{X}} - N_t\hat{\bar{\mathbf{U}}}_\pi\right)^T\left(N_t\hat{\bar{\mathbf{V}}}_\pi\right)^{-1}\left(N_t\bar{\mathbf{X}} - N_t\hat{\bar{\mathbf{U}}}_\pi\right)\right\} \\ &= N_t E_p\left\{\left(\bar{\mathbf{X}} - \hat{\bar{\mathbf{U}}}_\pi\right)^T\hat{\bar{\mathbf{V}}}_\pi^{-1}\left(\bar{\mathbf{X}} - \hat{\bar{\mathbf{U}}}_\pi\right)\right\} \\ &= N_t\left\{\left(\bar{\mathbf{X}} - \bar{\mathbf{U}}\right)^T\bar{\mathbf{V}}^{-1}\left(\bar{\mathbf{X}} - \bar{\mathbf{U}}\right) + o(1)\right\}. \end{aligned} \tag{A.2}$$

The result follows immediately from (A.1) and (A.2).

Lemma 1: Let $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ be scalars where $b_i > 0$ for all i. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be p-vectors. Then

$$\left(\sum_{i=1}^N a_i\mathbf{x}_i\right)^T\left(\sum_{i=1}^N b_i\mathbf{x}_i\mathbf{x}_i^T\right)^{-1}\left(\sum_{i=1}^N a_i\mathbf{x}_i\right) \le \sum_{i=1}^N a_i^2/b_i \tag{A.3}$$

provided the matrix inverse exists. Equality in (A.3) obtains if and only if

$$a_i b_i^{-1} = \boldsymbol{\lambda}^T\mathbf{x}_i \tag{A.4}$$

for all $i = 1, \ldots, n$ for some p-vector $\boldsymbol{\lambda}$.

Proof of Lemma 1

Let $b = \sum_{i=1}^n b_i$. Let X be a discrete random variable taking on the values a_i/b_i. Let $\mathbf{Y}$ be a discrete random variable taking on the values $\mathbf{x}_i$, for $i = 1, \ldots, n$. Let $P[Y = \mathbf{x}_i, X = a_i/b_i] = b_i/b$ for $i = 1, \ldots, n$. Write $M_1 \le M_2$ if $M_1 - M_2$ is negative semi-definite for any matrices M_1 and M_2. Theorem 1 of Tripathi (1999) states that for any random vectors $\mathbf{X}$ and $\mathbf{Y}$,

$$E\left[\mathbf{X}\mathbf{Y}^T\right]\left\{E\left[\mathbf{Y}\mathbf{Y}^T\right]\right\}^{-1}E\left[\mathbf{Y}\mathbf{X}^T\right] \le E\left[\mathbf{X}\mathbf{X}^T\right] \tag{A.5}$$

provided the matrix inverse exists. With my definition of X and $\mathbf{Y}$, (A.5) becomes

$$\sum_{i=1}^N b_i b^{-1} a_i b_i^{-1}\mathbf{x}_i^T\left\{\sum_{i=1}^N b_i b^{-1}\mathbf{x}_i\mathbf{x}_i^T\right\}^{-1}\sum_{i=1}^N b_i b^{-1} a_i b_i^{-1}\mathbf{x}_i \le \sum_{i=1}^N b_i b^{-1} a_i^2 b_i^{-2} \tag{A.6}$$

which leads directly to (A.3).Tripathi (1999) states that the equality is sharp if

$$\mathbf{X}^T\boldsymbol{\lambda}_1 + \mathbf{Y}^T\boldsymbol{\lambda}_2 = 0 \tag{A.7}$$

with probability 1 for some λ_1 of the same dimension as $\mathbf{X}$ and λ_2 of the same dimension as $\mathbf{Y}$. Here, (A.7) becomes

$$a_i b_i^{-1}\lambda_1 = \mathbf{x}_i^T\boldsymbol{\lambda}_2$$

for all i, which is equivalent to (A.4).

Proof of Theorem 2

Let $a_i = 1 - \pi_i$ and $b_i = \pi_i \sigma_i^{-2}$. From Lemma 1,

$$\begin{aligned} \text{AV} &= \sum_U (1-\pi_i)\,\mathbf{x}_i^T \left(\sum_U \pi_i \sigma_i^{-2} \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \sum_U (1-\pi_i)\,\mathbf{x}_i^T + \sum_U (1-\pi_i)\,\sigma_i^2 \\ &\le \sum_U (1-\pi_i)^2\, \pi_i^{-1} \sigma_i^2 + \sum_U (1-\pi_i)\,\sigma_i^2 \\ &= \sum_U (1-\pi_i)\,\pi_i^{-1} \sigma_i^2 \,(1 - \pi_i + \pi_i) \\ &= \sum_U \left(\pi_i^{-1} - 1\right) \sigma_i^2 \end{aligned}$$

with strict equality if and only if

$$\boldsymbol{\lambda}^T \mathbf{x}_i = a_i b_i^{-1} = \left(\pi_i^{-1} - 1\right) \sigma_i^2$$

for some vector $\boldsymbol{\lambda}$.

References

Breidt, F.J., Claeskens, G. and Opsomer, J.D. (2005). Model-assisted estimation for complex surveys using penalised splines. *Biometrika*, 92(4), 831-846.

Brewer, K.R.W. (1999). Cosmetic calibration with unequal probability sampling. *Survey Methodology*, 25, 2, 205-212. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1999002/article/4883-eng.pdf.

Brewer, K.R.W., Hanif, M. and Tam, S.M. (1988). How nearly can model-based prediction and design-based estimation be reconciled? *Journal of the American Statistical Association*, 83, 128-132.

Chambers, R., and Clark, R. (2012). *An Introduction to Model-Based Survey Sampling with Applications*. Oxford University Press: Oxford.

Godambe, V.P., and Joshi, V.M. (1965). Admissibility and Bayes estimation in sampling finite populations 1. *Annals of Mathematical Statistics*, 36, 1707-1722.

Hansen, M.H., Madow, W.G. and Tepping, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78, 776-793.

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

Kott, P.S. (1986). When a mean-of-ratios is the best linear unbiased estimator under a model. *The American Statistician*, 40(3), 202-204.

Lohr, S.L. (2010). *Sampling: Design and Analysis, 2nd edition*. Boston: Brooks-Cole.

Nedyalkova, D., and Tillé, Y. (2008). Optimal sampling and estimation strategies under the linear model. *Biometrika*, 95(3), 521-537.

Nedyalkova, D., and Tillé, Y. (2012). Bias-robustness and efficiency of model-based inference in survey sampling. *Statistica Sinica*, 22(2), 777-794.

Royall, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57(2), 377-387.

Royall, R.M. (1992). Robustness and optimal design under prediction models for finite populations. *Survey Methodology*, 18, 2, 179-185. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1992002/article/14488-eng.pdf.

Royall, R.M., and Herson, J. (1973). Robust estimation in finite populations I. *Journal of the American Statistical Association*, 68(344), 880-889.

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

Scott, A.J., Brewer, K.R.W. and Ho, E.W.H. (1978). Finite population sampling and robust estimation. *Journal of the American Statistical Association*, 73(362), 359-361.

Steel, D.G., and Clark, R.G. (2014). Potential gains from using unit level cost information in a model-assisted framework. *Survey Methodology*, 40, 2, 231-242. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2014002/article/14110-eng.pdf.

Tam, S.M. (1988). Asymptotically design-unbiased predictors in survey sampling. *Biometrika*, 75(1), 175-177.

Tillé, Y. (2006). *Sampling Algorithms*. New York: Springer.

Tillé, Y., and Matei, A. (2016). *Sampling: Survey Sampling*. R package version 2.8. https://CRAN.R-project.org/package=sampling.

Tripathi, G. (1999). A matrix extension of the Cauchy-Schwarz inequality. *Economics Letters*, 63(1), 1-3.

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

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

Welsh, A.H., and Wiens, D.P. (2013). Robust model-based sampling designs. *Statistics and Computing*, 23(6), 689-701, November. https://doi.org/10.1007/s11222-012-9339-3.

Survey Methodology, June 2020
Vol. 46, No. 1, pp. 93-119
Statistics Canada, Catalogue No. 12-001-X

Considering interviewer and design effects when planning sample sizes

Stefan Zins and Jan Pablo Burgard[1]

Abstract

Selecting the right sample size is central to ensure the quality of a survey. The state of the art is to account for complex sampling designs by calculating effective sample sizes. These effective sample sizes are determined using the design effect of central variables of interest. However, in face-to-face surveys empirical estimates of design effects are often suspected to be conflated with the impact of the interviewers. This typically leads to an over-estimation of design effects and consequently risks misallocating resources towards a higher sample size instead of using more interviewers or improving measurement accuracy. Therefore, we propose a corrected design effect that separates the interviewer effect from the effects of the sampling design on the sampling variance. The ability to estimate the corrected design effect is tested using a simulation study. In this respect, we address disentangling cluster and interviewer variance. Corrected design effects are estimated for data from the European Social Survey (ESS) round 6 and compared with conventional design effect estimates. Furthermore, we show that for some countries in the ESS round 6 the estimates of conventional design effect are indeed strongly inflated by interviewer effects.

Key Words: Design effect; Interviewer effect; Multilevel model; Sample size; European Social Survey (ESS).

1 Introduction

Determining the sample size of a survey can be very demanding. The complexity of the task is often exacerbated by a lack of information and data on which to plan the survey. That is why survey planners seek to reduce the complexity of the problem using simplifications and statistical models. One such approach is to use the so-called *design effect* to select a sample size. The design effect is then defined as the ratio between the variance of an estimator under the sampling design of the planned survey and the variance of the same estimator under a simple random sample design. As such, the design effect is a property of an estimation strategy, i.e., a sampling design and an estimator (Chaudhuri and Stenger, 2005, page 4), not of the survey. The weighted sample mean of a single variable is usually used as a reference estimator. However, for reasons of simplification, if we speak in the following of the design effect of a sampling design, then we do this always with respect to the sampling variance of a weighted sample mean.

To plan the sample size, an effective sample size target can be set, meaning that the planned sample size divided by the planned design effect should be above a certain value. The effective sample size of a sampling design is the simple random sample equivalent of its sample size, in terms of efficiency, i.e., if a sampling design has an effective sample size of 1,000, then its sampling variance is equal to that of a simple random sample of size 1,000.

Ideally, a survey planner designs a survey with a specific analysis or hypotheses test in mind and formulates their opinion about tolerable sampling error levels or type II error probabilities. These opinions should be based on two things. First, some level of experience with the substantial research question, and

[1] Stefan Zins, Institute for Employment Research (IAB) of the German Federal Employment Agency (BA), Regensburger Strasse 104, D-90478 Nürnberg. E-mail: st.zins@gmail.com; Jan Pablo Burgard, RIFOSS - Research Institute for Official and Survey Statistics, Trier University, D-54286 Trier.

second, on assumptions over target population parameters necessary for sampling error planning and power calculations. Assumptions about target population parameters can stem from previous rounds of a survey, or be based on data collected during the field test for the survey. Power calculations and sampling error planning are much less complex and require less information about the target population if done under the assumption of a simple random sampling design. That is why most methods addressing sample size planning found in textbooks are suited for determining an effective sample size. The effect of complex sampling is then factored in by multiplying the planned effective sample size with a planned design effect. Determining a design effect can thus be separated from selecting an effective sample size. For example, if a simple random sample of size 1,000 ensures the following: The sampling error of an estimator does not exceed a given value with a probability of 95%, or that the power of a statistical test is 80%, that is, the probability of rejecting a null hypothesis in case the alternative is true should be 80% (Ellis, 2010, Chapter 3). Then multiplying 1,000 by the assumed design effect of the a study will give the survey planner the required net sample size to achieve set precision targets.

The decision on an effective sample size also has to reflect a certain trade-off between the cost of the survey and the precision of survey estimates. Regarding this trade-off, the survey planner should, for example, consider what the consequences are if a type II error is committed, i.e., if a null hypothesis is not rejected even though the alternative hypothesis is true.

For surveys that are primarily intended for secondary analysis, i.e., they provide data to the research community with no single application in mind, like the European Social Survey (ESS) or the European Value Study (EVS), the decision on an effective sample size cannot be planned for a single research question or hypothesis test. For that reason, the ESS uses an average effective sample size. This means that ESS sample designs are planned such that the average design effect for a set of items from the ESS core questionnaire should have a certain value. The planned average design effect is multiplied by the required average effective sample size to calculate the planned net sample size. The net sample size is the sample size after unit-nonresponse, i.e., the number of completed interviews. To plan the gross sample size – that is, the sample size before unit-nonresponse – the net sample size is divided by the product of the assumed response rate and eligibility rate. The eligibility rate is the fraction of sampled persons that belong to the target population, which can be lower than 100% because of sampling frame imperfections.

However, design effects can still be difficult to quantify, given the complexity of the sampling design. Hence, to reduce complexity, statistical models for survey data are used to approximate the design effect. Such models commonly try to incorporate the effect of cluster sampling, which can have a large effect on the sampling variance of estimates. Clusters can be spatial areas like settlements, organizational units like municipalities, or institutions such as hospitals and schools. They are either used as so-called *Primary Sampling Units* (PSUs), which are selected first and then an additional sampling takes place within them, or they are surveyed in their entirety. For example, the German ESS round 6 (ESS6) sampling design has two sampling stages. The PSUs are municipalities, and the secondary sampling units are persons registered within the municipalities. Variables of interest can often not be considered as identically distributed over all clusters in the population. In fact, it can be assumed that respondents within the same cluster are usually more similar to one another than those belonging to a different cluster. Kish (1965), page 162, gives the following formula for a design effect due to clustering:

$$\text{deff} = 1 + (b - 1)\,\rho. \tag{1.1}$$

This design effect deff consists of two parameters, b is typically an average cluster size in terms of realized respondents, and ρ, the intra-cluster correlation coefficient, which is a measure for the homogeneity of the measurements of a variable within the same cluster. ρ can be defined using variance decomposition as the between-cluster variance divided by the sum of the within-cluster and between-cluster variances. The higher the variance between the clusters the higher ρ will be.

To use deff when selecting a sample size, assumptions have to be made about the unknown parameter ρ. The cluster size b does not depend on the measured variable and can be influenced by the survey planner. For ρ, data from previous surveys can be used to formulate the necessary assumption. Especially for repeated cross-sectional surveys, their accumulated data is of great help in planning the sampling design for the next implementation of the survey.

Lynn, Häder, Gabler and Laaksonen (2007) describe how predicted design effects are used by the ESS to plan sample sizes that achieve a certain average effective sample size under a given sampling design. For recent rounds of the ESS, the prediction of the design effect and its components was informed by estimates of these statistics based on data from the preceding ESS rounds (The ESS Sampling Expert Panel, 2016).

An important factor that can also introduce homogeneity to measurements in face-to-face surveys is the interviewer. Embedded in the Total Survey Error (TSE) framework (Groves, 2009), different mechanisms have been described for how an interviewer can influence survey measurements. Similar to cluster sampling, interviewers have long been identified as a source of dependent measurements (Kish, 1965, page 522, Kish, 1962),with interviewers introducing homogeneity through measurement errors and selection effects, rather than the homogeneity of clusters that is intrinsic to the population. West and Blom (2017) give an overview of the research on interviewer effects. They detail how interviewer tasks like generating and/or applying sampling frames, making contact, and gaining cooperation and consent can have a selection effect on the recruitment of respondents. West and Blom (2017) also outline evidence that interviewers conducting measurements, making observations and finally recording the gathered information can introduce measurement and processing errors into the data that is used for analysis. For an overview of other sources of variance in surveys, we refer to the TSE framework as described, e.g., by Groves and Lyberg (2010) and Biemer (2010).

Analysis of interviewer effects using ESS data from different countries and years showed that this effect can be considerable (Beullens and Loosveldt, 2016). Such findings raise a question: To what extent ρ in equation (1.1) is driven by intra-cluster correlation, rather than intra-interviewer correlation? Schnell and Kreuter (2005) show that the interviewer effect can be higher than the cluster effect, even for variables where a strong spatial correlation can be assumed. Consequently, the estimated design effect for face-to-face surveys is typically conflated with the interviewer effect. Hence, the design effect is systematically over-estimated in face-to-face surveys. This might pose a problem to surveys that predict design effects using historical data to plan sample sizes, as there is a risk of misallocating funds. A survey planner could try to offset an increase in the predicted design effect by increasing the sample size to hold the effective sample size constant. If the driving factor inflating the predicted design effect is the interviewer effect, funds could be more effectively allocated by hiring additional interviewers and/or training them better to improve measurement accuracy and reduce selection effects.

The novel part of the presented approach is that the proposed method allows for estimating a corrected design effect that is not conflated with the interviewer effects. With the proposed corrected design effect, the survey planner is able to make evidence-based decisions on changes in the sampling design, such as sample size and number of PSUs, and/or about the deployment of interviewers.

The article is structured as follows: Section 2 introduces the framework for describing the effects of the sampling design and the interviewer. The framework follows the model based justification of the design effect as outlined by Gabler, Häder and Lahiri (1999) and the introduction of an interviewer effect to this framework by Gabler and Lahiri (2009). The measurement models used to describe the observed data follow a multilevel structure. The influence of multi-stage or cluster sampling, and that of interviewers on the observed data, is modeled with the help of random effects that imply a certain variance-covariance structure. This approach allows for a factorization of the overall effect into separate sampling and interviewer effects. This separation is essential when addressing effects separately in order to control for them.

In Section 3, the sampling and interviewer effects described in Section 2 are estimated for ESS6 data with the help of multilevel models. First, we present the results from a simulation study conducted to assess the possibility of disentangling cluster and interviewer variances for the observed PSU-interviewer structure in the ESS6 data. Afterwards, we evaluate the applicability of the different measurement models for a selected set of ESS variables. The selected models are used to estimate the variances of different random effects in multilevel models, which are in turn used for estimating the intra-PSU and intra-interviewer correlation.

In Section 4, we present our conclusions and give recommendations for survey planners based on both our theoretical work in Section 2 and the empirical findings in Section 3. We then point to possible future research to adapt our relatively simplistic measurements models to better reflect complex sampling designs and the heterogeneity of interviewers.

2 Interviewer and design effects

We define a sample as a set of n distinct respondents, which we denote as $s = \{1, \ldots, n\}$, with $n \in \mathbb{N}$. For the k^{th} respondent our variable of interest y is a real valued variable, where y_k is the observation of this variable for the k^{th} respondent in our sample s. The observed data is given by $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$. We associate survey weights with every respondent in the sample, given by $\mathbf{w} = (w_1, \ldots, w_n)^{\mathrm{T}}$, where w_k is the weight of the k^{th} respondent and $w_k > 0$, for all $k \in s$.

We consider the weighted sample mean of $\mathbf{y}$ as our estimator, given by

$$\bar{y}(\mathbf{w}) = \frac{\mathbf{w}^{\mathrm{T}}\mathbf{y}}{\mathbf{w}^{\mathrm{T}}\mathbb{I}_n} = \frac{\sum_{k \in s} w_k y_k}{\sum_{k \in s} w_k}, \tag{2.1}$$

where $\mathbb{I}_n$ is a column vector of ones of length n. We focus on one estimator of interest, $\bar{y}(\mathbf{w})$, as it is the most common choice for describing interviewer and design effects (Kish, 1965, Section 8.1, Kish, 1962; Särndal, Swensson and Wretman, 1992, page 53). This choice enables us to use an established framework (Gabler et al., 1999) and produce formulas that are recognizable to readers that are already somewhat familiar with the topic. However, design effects of other estimators have been studied, notably, Lohr (2014), derives design effects for estimators of regression coefficients and Fischer, West, Elliott and Kreuter (2018), describe the impact of interviewer effects on the estimation of regression coefficients.

In the following, the variance of $\bar{y}(\mathbf{w})$ is derived under different measurement models for y. The different models serve to distinguish between complex and simple sampling designs, as well as when there is and is not an interviewer effect. It should be noted that the model based variance of estimator $\bar{y}(\mathbf{w})$, which we use, is, in general, not the same as its design based variances, i.e., the variance of $\bar{y}(\mathbf{w})$ under a given sampling design (Särndal et al., 1992, page 492). Design based variances can be very complex and thus difficult to display in an accessible fashion, especially for multi-stage sampling. The model based approach reduces complexity while retaining the essential property of the complex sampling designs that we study, the cluster effect of multi-stage sampling. It also makes it possible to easily integrate cluster and interviewer effect into a common framework.

2.1 Simple random sampling without an interviewer effect

To model simple random sampling in the absence of an interviewer effect, i.e., without intra-PSU and intra-interviewer correlation, we assume the following measurement model (M_0)

$$y_k = \mu_k + \mathrm{e}_k, \tag{M_0}$$

where μ_k is the value of y for the k^{th} respondent and e_k is the measurement error. The measurement errors e_k for all $k \in s$ are independent and identically distributed (iid) random variables with a variance-covariance structure of

$$\mathrm{Cov}_{M_0}(e_k, e_l) = \begin{cases} \sigma^2, & \text{if } k = l \\ 0, & \text{else} \end{cases}, \tag{2.2}$$

where σ is a real value parameter greater than zero. Under model (M_0), the variance of $\bar{y}(\mathbb{I}_n)$ is given by $V_{M_0}(\bar{y}(\mathbb{I}_n)) = \sigma^2/n$. This variance can be interpreted as the variance of the unweighted sample mean under simple random sampling with replacement (Särndal et al., 1992, page 73). Simple random sampling with estimator $\bar{y}(\mathbb{I}_n)$ typically serves as a reference estimation strategy, which is compared with more complex sampling designs and estimators.

2.2 Simple random sampling with an interviewer effect

Next, we introduce interviewer variance into our measurement model for y. Each respondent is interviewed by one and only one interviewer. There are $R \in \mathbb{N}_{>0}$, interviewers that conduct the interviews of all n respondents. We denote $s_i \subset s$ as the set of all respondents that are interviewed by the i^{th} interviewer and $\mathcal{R} = \{1, \ldots, R\}$ as the set of all interviewers. The workload of the i^{th} interviewer is given by n_i, $\mathbf{n}_I = (n_1, \ldots, n_R)^{\top}$ is the vector of interviewer workloads and $\sum_{i=1}^{R} n_i = n$. Under measurement model (M_1), which follows the explanations of Särndal et al. (1992), page 623, the observed values of y for $k \in s_i$ are described as

$$y_{ik} = \mu_k + \mathfrak{I}_i + \mathbb{e}_{ik}, \tag{M_1}$$

with $\mathfrak{I}_i$ being the interviewer effect associated with all measurements conducted for respondents $k \in s_i$. $\mathbb{e}_{ik}$ represents the random error due to sources other than the interviewer. All $\mathbb{e}_{ik}$ for $i \in \mathcal{R}$ and $k \in s$ are iid random variables with zero mean and variance σ_e^2. $\mathfrak{I}_1, \ldots, \mathfrak{I}_R$ are iid random variables with zero mean and variance σ_I^2, which we call interviewer variance, and they are independent of $\mathbb{e}_{ik}$ for all $i \in \mathcal{R}$ and $k \in s$. Särndal et al. (1992) interprets model (M_1) as a random assignment of interviewers to a pre-defined partition of the sample s into R disjoint subsets s_i, $i = 1, \ldots, R$. These subsets could correspond to different geographical areas where the survey is conducted and the interviewers are then randomly allocated to them. In practice, in many surveys fieldwork agencies assign interviewers to geographical areas based on experience and proximity. As this process is not necessarily observable by the researcher estimating the design effect, we assume a random allocation of interviewers to the PSUs. This can be seen as the recruitment of interviewers from an infinite, or very large, pool of possible interviewers.

If we define the random part in y_{ik} as $\varepsilon_{ik} = \mathfrak{I}_i + \mathbb{e}_{ik}$, then the variance-covariance structure of y_{ik} under model (M_1) is given by

$$\operatorname{Cov}_{M_1}(\varepsilon_{ik}, \varepsilon_{jl}) = \begin{cases} \sigma^2, & \text{if } i = j,\ k = l \\ \rho_I \sigma^2, & \text{if } i = j,\ k \neq l, \\ 0, & \text{else} \end{cases} \tag{2.3}$$

where $\sigma_I^2 + \sigma_e^2 = \sigma^2$ and $\rho_I = \frac{\sigma_I^2}{\sigma^2}$ is the correlation between two different observations of y made by the same interviewer. To derive the variance of $\bar{y}(\mathbf{w})$ under model (M_1), we first determine the variance of $\sum_{i \in \mathcal{R}} \sum_{k \in s_i} w_{ik} y_{ik}$, where w_{ik} and y_{ik} are the survey weight and the observation for respondent $k \in s_i$, respectively. Thus we have

$$\begin{aligned}
\operatorname{Var}_{M_1}\left(\sum_{i \in \mathcal{R}} \sum_{k \in s_i} w_{ik} y_{ik}\right) &= \sigma^2 \left(\sum_{i \in \mathcal{R}} \sum_{k \in s_i} w_{ik}^2 + \rho_I \sum_{i \in \mathcal{R}} \sum_{k \in s_i} \sum_{\substack{l \in s_i \\ l \neq k}} w_{ik} w_{il}\right) \\
&= \sigma^2 \left(\sum_{i \in \mathcal{R}} \sum_{k \in s_i} w_{ik}^2 + \rho_I \left[\sum_{i \in \mathcal{R}} \left(\sum_{k \in s_i} w_{ik}\right)^2 - \sum_{i \in \mathcal{R}} \sum_{k \in s_i} w_{ik}^2\right]\right) \\
&= \sigma^2 \sum_{i \in \mathcal{R}} \sum_{k \in s_i} w_{ik}^2 \left(1 + \rho_I \left[\frac{\sum_{i \in \mathcal{R}} \left(\sum_{k \in s_i} w_{ik}\right)^2}{\sum_{i \in \mathcal{R}} \sum_{k \in s_i} w_{ik}^2} - 1\right]\right),
\end{aligned}$$

from which follows

$$\operatorname{Var}_{M_1}(\bar{y}(\mathbf{w})) = \frac{\sigma^2 \sum_{i \in \mathcal{R}} \sum_{k \in s_i} w_{ik}^2}{\left(\sum_{i \in \mathcal{R}} \sum_{k \in s_i} w_{ik}\right)^2} \left(1 + \rho_I \left[\frac{\sum_{i \in \mathcal{R}} \left(\sum_{k \in s_i} w_{ik}\right)^2}{\sum_{i \in \mathcal{R}} \sum_{k \in s_i} w_{ik}^2} - 1\right]\right). \tag{2.4}$$

2.3 Multi-stage sampling with an interviewer effect

We consider a two-stage sampling design, where first PSUs are selected, and at the second stage respondents are selected from within the sampled PSUs. PSUs are the clustering units and we will treat the terms cluster and PSU as interchangeable. The sample of PSUs is denoted $\mathcal{K} = \{1, \ldots, K\}$, with $K > 1$. Each respondent belongs to one PSU and one PSU only. Let $s_q \subset s$ be the set of all respondents belonging to the q^{th} PSU, n_q be the number of respondents observed within the q^{th} PSU, $\mathbf{n}_C = (n_1, \ldots, n_K)^{\top}$ the vector of cluster sizes, and $\sum_{q \in \mathcal{K}} n_q = n.$ Again, each respondent is interviewed by one interviewer and one interviewer only. Interviewers can work across PSUs and PSUs can be visited by multiple interviewers. Although interviewers might concentrate their work in a particular region, these regions are usually composed of multiple PSUs and interviewers do not work exclusively in one PSU only. This situation is frequently found in face-to-face surveys across Europe, e.g., in the ESS or EVS. Table 3.1 in Section 3.1 gives an overview on the level of interpenetration between PSUs and interviewer for countries that use a multi-stage sampling design in ESS6. Interpenetration between PSUs and interviewer can be observed across all ESS rounds for countries that use multi-stage sampling design.

We now introduce measurement model (M_2), which incorporates both cluster and interviewer variance into the observed values of $y.$ For $k \in s_{qi} = s_q \cap s_i$ we model observations of y as

$$y_{qik} = \mu_k + \mathfrak{C}_q + \mathfrak{I}_i + \mathrm{e}_{qik}, \tag{M_2}$$

with $\mathfrak{C}_q$ defined as a random variable with mean zero and variance σ_C^2, which we call PSU variance, common to all respondents in PSU q. $\mathfrak{C}_1, \ldots, \mathfrak{C}_K$ are iid random variables and are independent of e_{qik}

and $\mathfrak{I}_i$ for all $i \in \mathcal{R}, q \in \mathcal{K},$ and $k \in s_{qi}$. $\mathfrak{C}_q$ introduces a certain degree of similarity between respondents from the same PSU. It allows for a permanent random effect of the PSU on the measurement of y, for the k^{th} respondent, causing it to deviate from μ_k (Chambers and Skinner, 2003, page 201).

To establish the effect of sampling and interviewers on $\bar{y}(\mathbf{w})$, we define the random part of y_{qik} as $\varepsilon_{qik} = \mathfrak{C}_q + \mathfrak{I}_i + \mathbb{e}_{qik}$, which has the following variance-covariance structure

$$\operatorname{Cov}_{M_2}\left(\varepsilon_{qik}, \varepsilon_{pjl}\right) = \begin{cases} \sigma^2, & \text{if } q = p,\ i = j,\ k = l \\ \rho_C \sigma^2, & \text{if } q = p,\ i \neq j,\ k \neq l \\ \rho_I \sigma^2, & \text{if } q \neq p,\ i = j,\ k \neq l \\ (\rho_I + \rho_C)\sigma^2, & \text{if } q = p,\ i = j,\ k \neq l \\ 0, & \text{else} \end{cases}, \tag{2.5}$$

where $\sigma_C^2 + \sigma_I^2 + \sigma_e^2 = \sigma^2$ and $\rho_C = \frac{\sigma_C^2}{\sigma^2}$ is the correlation between observation from the same PSU. The variance-covariance structure of ε_{qik} implies that the measurements of y are correlated if they are made within the same PSU or the same interviewer. Further, measurements of y are more homogeneous if they are made by the same interviewer within the same PSU. Model (M_2) represents a generalization of model M_4 of Gabler and Lahiri (2009), by removing the restriction that no interviewer works in more than one PSU.

The variance of $\bar{y}(\mathbf{w})$ under model (M_2) is given by

$$\operatorname{Var}_{M_2}(\bar{y}(\mathbf{w})) = \frac{\sigma^2 \sum_{q \in \mathcal{K}} \sum_{i \in \mathcal{R}} \sum_{k \in s_{qi}} w_{qik}^2}{\left(\sum_{q \in \mathcal{K}} \sum_{i \in \mathcal{R}} \sum_{k \in s_{qi}} w_{qik}\right)^2} \left(1 + \rho_I\left[\bar{m}_I(\mathbf{w}) - 1\right] + \rho_C\left[\bar{m}_C(\mathbf{w}) - 1\right]\right), \tag{2.6}$$

where w_{qik} and y_{qik} are the survey weight and the observation for respondent $k \in s_{qi}$, respectively, and

$$\bar{m}_I(\mathbf{w}) = \frac{\sum_{i \in \mathcal{R}} \left(\sum_{q \in \mathcal{K}} \sum_{k \in s_{qi}} w_{qik}\right)^2}{\sum_{q \in \mathcal{K}} \sum_{i \in \mathcal{R}} \sum_{k \in s_{qi}} w_{qik}^2} \quad \text{and} \quad \bar{m}_C(\mathbf{w}) = \frac{\sum_{q \in \mathcal{K}} \left(\sum_{i \in \mathcal{R}} \sum_{k \in s_{qi}} w_{qik}\right)^2}{\sum_{q \in \mathcal{K}} \sum_{i \in \mathcal{R}} \sum_{k \in s_{qi}} w_{qik}^2}.$$

We can alter model (M_2) to allow for a PSU interviewer interaction effect, meaning that the covariance between the observations made by the same interviewer within the same PSU is not equal to the sum of the intra-PSU and intra-interviewer covariance. We call this measurement model (M_{2*}) and for $k \in s_{qi}$ the observation of y is modeled as

$$y_{qik} = \mu_k + \mathfrak{C}_q + \mathfrak{I}_i + \mathfrak{D}_{qi} + \mathbb{e}_{qik}, \tag{M_{2*}}$$

with $\mathfrak{D}_{qi}$ as a random variable with mean zero and variance σ_{IC}^2 common to all respondents in PSU q that were interviewed by interviewer i. All $\mathfrak{D}_{qi}$ for $q \in \mathcal{K}$ and $i \in \mathcal{R}$ are iid random variables and are independent of $\mathbb{e}_{qik}, \mathfrak{I}_i, \mathfrak{C}_q$ for all $q \in \mathcal{K}, i \in \mathcal{R},$ and $k \in s_{qi}$. Random effect $\mathfrak{D}_{qi}$ introduces some additional correlation between observations made by the same interviewer within the same PSU, which cannot be explained by the separate PSU and interviewer variances.

For $k \neq l$ and $\varepsilon_{qik} = \mathfrak{C}_q + \mathfrak{I}_i + \mathfrak{D}_{qi} + \mathbb{e}_{qik}$ we have under model (M_{2*}) $\operatorname{Cov}_{M_{2*}}(\varepsilon_{qik}, \varepsilon_{qil}) = (\rho_I + \rho_C + \rho_{IC})\,\sigma^2$. Thus, we can write the variance of $\bar{y}(\mathbf{w})$ under model (M_{2*}) as

$$\begin{aligned}\operatorname{Var}_{M_{2*}}(\bar{y}(\mathbf{w})) &= \frac{\sigma^2 \sum_{q \in \mathcal{K}} \sum_{i \in \mathcal{R}} \sum_{k \in s_{qi}} w_{qik}^2}{\left(\sum_{q \in \mathcal{K}} \sum_{i \in \mathcal{R}} \sum_{k \in s_{qi}} w_{qik}\right)^2} \\ &\quad (1 + \rho_I [\bar{m}_I(\mathbf{w}) - 1] + \rho_C [\bar{m}_C(\mathbf{w}) - 1] + \rho_{IC} [\bar{m}_{IC}(\mathbf{w}) - 1]), \end{aligned} \tag{2.7}$$

where

$$\bar{m}_{IC}(\mathbf{w}) = \frac{\sum_{q \in \mathcal{K}} \sum_{i \in \mathcal{R}} \left(\sum_{k \in s_{qi}} w_{qik}\right)^2}{\sum_{q \in \mathcal{K}} \sum_{i \in \mathcal{R}} \sum_{k \in s_{qi}} w_{qik}^2}.$$

2.4 Survey effect

After we establish the variance of $\bar{y}(\mathbf{w})$ under the different measurement models, we can define the effect associated with complex sampling and interviewers. We will refer to this effect as the *survey effect*, which we define as

$$\operatorname{eff}_{ab}(\mathbf{w}) = \frac{\operatorname{Var}_{M_a}(\bar{y}(\mathbf{w}))}{\operatorname{Var}_{M_b}(\bar{y}(\mathbf{w}))}, \tag{2.8}$$

where M_a is the measurement model assumed for our survey of interest and M_b is the reference model. We use the term *survey effect* to distinguish $\operatorname{eff}_{ab}(\mathbf{w})$ from design and interviewer effect, as $\operatorname{eff}_{ab}(\mathbf{w})$ incorporates both effects. Other sources of variance, as described in the TSE framework, are not considered. Consequently, we will use the term *survey design* for the combination of a sampling design and interviewer workplan.

The survey effect associated with measurement model (M_2), is given by

$$\begin{aligned}\operatorname{eff}_{20}(\mathbf{w}) &= \frac{\operatorname{Var}_{M_2}(\bar{y}_w)}{\operatorname{Var}_{M_0}(\bar{y})} \\ &= \operatorname{eff}_w(\mathbf{w})\,(1 + \rho_I [\bar{m}_I(\mathbf{w}) - 1] + \rho_C [\bar{m}_C(\mathbf{w}) - 1]), \end{aligned} \tag{2.9}$$

where

$$\operatorname{eff}_w(\mathbf{w}) = \frac{n \sum_{k \in s} w_k^2}{\left(\sum_{k \in s} w_k\right)^2} \geq 1.$$

Factor $\mathrm{eff}_w(\mathbf{w})$ does not depend on the measurement model and can be interpreted as a measure for the variance of the weights $\mathbf{w}$. If we write the variance of the weights as $\sigma_{\mathbf{w}}^2 = 1/n \sum_{k \in s} w_k^2 - \bar{w}^2$, with $\bar{w} = 1/n \sum_{k \in s} w_k$, this relationship becomes more clear, as $\mathrm{eff}_w(\mathbf{w}) = \mathrm{CV}_{\mathbf{w}}^2 + 1$, with $\mathrm{CV}_{\mathbf{w}} = \sigma_{\mathbf{w}} / \bar{w}$ as the coefficient of variation of the survey weights. If the weights are all equal, then $\mathrm{CV}_{\mathbf{w}} = 0$ and $\mathrm{eff}_w(\mathbf{w})$ becomes 1. Terms $\bar{m}_I(\mathbf{w})$ and $\bar{m}_C(\mathbf{w})$ can be seen as measures for the average workload of the interviewers and the PSU size, respectively. If all weights are equal, $\bar{m}_I(\mathbf{w})$ has the value $\bar{m}_I(\mathbb{I}_n) = \sum_{i \in \mathcal{R}} n_i^2 / n$. Furthermore, if all interviewers have the exact same workload, i.e., $n_i = n/R$ for $i = 1, \ldots, R$, we have $\bar{m}_I(\mathbb{I}_n) = n/R$. $\bar{m}_C(\mathbf{w})$ has similar properties.

Following Gabler et al. (1999) and Gabler and Lahiri (2009) we can give the following upper bound for the survey effect.

Result 1.

$$\mathrm{eff}_{20}^*(\mathbf{w}) \leq \mathrm{eff}_w(\mathbf{w}) \mathrm{eff}_{20}^*(\mathbb{I}_n),$$

where eff_{20}^* is the survey effect under the condition that $n_i = n/R$ for all $i \in \mathcal{R}$ and $n_q = n/K$ for all $q \in \mathcal{K}$. The upper bound of $\mathrm{eff}_{20}^*(\mathbf{w})$, given in Result 1, follows from $\bar{m}_I(\mathbf{w}) \leq n/R$ if $n_i = n/R$ for all $i \in \mathcal{R}$ (Gabler et al., 1999). The proof is given in the Appendix. For $\bar{m}_C(\mathbf{w})$ an analogous result holds. It should be noted that, in general, we do not have

$$\mathrm{eff}_{20}(\mathbf{w}) \geq \mathrm{eff}_{20}(\mathbb{I}_n) = 1 + \rho_I \left[\sum_{i \in \mathcal{R}} \frac{n_i^2}{n} - 1 \right] + \rho_C \left[\sum_{q \in \mathcal{K}} \frac{n_q^2}{n} - 1 \right]. \tag{2.10}$$

That is, we cannot say that the survey effect is greater or equal to the survey effect of an equally weighted design. If the weights have the same relative frequency distribution across all sets s_{qi} inequality (2.10) holds (Gabler and Lahiri, 2009), i.e., if we have

$$n_{qig} = \frac{n_{qi}}{n} n_g, \quad g = 1, \ldots, G, \tag{2.11}$$

where G is the number of unique values in $\mathbf{w}$, n_g the frequency of the g^{th} weighting value, and n_{qig} the frequency of the g^{th} weighting value for respondents interviewed by the i^{th} interviewer in the q^{th} PSU.

We can, however, give a lower bound to $\mathrm{eff}_{20}(\mathbf{w})$. Using the same argument that Gabler and Lahiri (2009) give in the proof of their Result 6, we get

$$\mathrm{eff}_{20}(\mathbf{w}) \geq \left(1 + \rho_I \left[\frac{n}{R} - 1 \right] + \rho_C \left[\frac{n}{K} - 1 \right] \right). \tag{2.12}$$

With the right-hand side of inequality (2.12) an easy to calculate minimum of $\mathrm{eff}_{20}(\mathbf{w})$ is given, which does not depend on the weights, the distribution of interviewer workloads, or the PSU sizes. This gives some valuable guidance at the planning stage of a survey design, as the planned survey effect of the survey should be at least as high as $\mathrm{eff}_{20}^*(\mathbb{I}_n)$. The practical utility of the upper bound in Result 1 is somewhat limited by strong assumptions about $\mathbf{n}_I$ and $\mathbf{n}_C$. The further the values of $\mathbf{n}_I$ and $\mathbf{n}_C$ deviate from the one point distribution of interviewer workloads and PSU sizes, the less this bound should serve as a guide. To give survey planners a less complex statistic to plan the value of $\bar{m}_I(\mathbf{w})$, Lynn and Gabler (2004) proposed using

$$\bar{m}'_I(\mathbf{w}) = \frac{H_{\mathbf{n}_I}}{H_{\mathbf{w}}}, \tag{2.13}$$

as a predictor for $\bar{m}_I(\mathbf{w})$, where $H_{\mathbf{n}_I} = \sum_{i \in \mathcal{R}} (n_i / n)^2$ is the Herfindahl index for the interviewer workload, a concentration measure, with $1/R \leq H_{\mathbf{n}_I} \leq 1$ (Fahrmeir, Heumann, Künstler, Pigeot and Tutz, 1997, page 83). $H_{\mathbf{n}_I} = 1$ corresponds to $R = 1$ and $H_{\mathbf{n}_I} = 1/R$ corresponds to $n_i = n/R$ for all $i \in \mathcal{R}$. $H_{\mathbf{w}} = \sum_{k \in s} \left(w_k / \sum_{k \in s} w_k \right)^2$ is the Herfindahl index for the weights. If equation (2.11) holds, we have $\bar{m}_I(\mathbf{w}) = \bar{m}'_I(\mathbf{w})$, but for most surveys this will not apply. For that reason, Lynn and Gabler (2004) suggested looking at $\operatorname{Cov}(w_{qik}, n_i)$, the covariance between the weights and interviewer workloads. The closer $\operatorname{Cov}(w_{qik}, n_i)$ is to zero the smaller the distance between $\bar{m}_I(\mathbf{w})$ and $\bar{m}'_I(\mathbf{w})$. Planning a survey with assumed values for $H_{\mathbf{n}_I}$ and $H_{\mathbf{w}}$ should be easier than with exact values of $\mathbf{n}_I$ and $\mathbf{w}$. Finding reasonable values for $H_{\mathbf{n}_I}$ and $H_{\mathbf{w}}$ could be guided by comparing these values from surveys with similar survey designs. Under equation (2.11) the findings are analogous for $\bar{m}_C(\mathbf{w})$.

It should be noted that we can also write $\operatorname{eff}_w(\mathbf{w})$ as

$$\operatorname{eff}_w(\mathbf{w}) = H_{\mathbf{w}} n. \tag{2.14}$$

The expression of $\operatorname{eff}_w(\mathbf{w})$ in equation (2.14) might also be useful at the planning stage of a survey, showing that it is possible to plan with a certain weight concentration, instead of specific values for $\mathbf{w}$.

Giving a general close upper bound for $\operatorname{eff}_{20}(\mathbf{w})$ is difficult if there are no restrictions on the values of $\mathbf{n}_I$, $\mathbf{n}_C$ and $\mathbf{w}$. However, survey weights are usually scaled to either the sample or the population size and it is not uncommon for them to be bounded. For example, the ESS provides weights to its users that are greater than zero and smaller or equal to 4 and scales them to the sample size (ESS, 2014c, 2014b). If $a \leq w_k \leq b$ for all $k \in s$ with $b < \infty$ and $a > 0$, then with a given value for $\mathbf{n}_I$ (or $\mathbf{n}_C$) upper limits of $\bar{m}_I(\mathbf{w})$, (or $\bar{m}_C(\mathbf{w})$) can be found, by solving a linear optimization problem. An upper limit for $\operatorname{eff}_w(\mathbf{w})$ can be deduced for given values of a and b, as shown in equation (A.5) in the Appendix.

The obtained upper bound of $\operatorname{eff}_{20}(\mathbf{w})$ will correspond to weight distributions with a very high concentration, i.e., a maximal number of the highest possible weights. However, adjusting the constraints of the linear optimization problem, based on the weight distribution of surveys with comparable sampling designs, can help to find bounds that are of higher practical relevance. (See Appendix for the formulation of this linear program.)

2.5 Corrected design effect

Now that we have established the survey effect of a survey design, we propose a new type of survey effect that we call *corrected design effect*. This statistic aims at quantifying the marginal effect of a complex survey design if an interviewer effect is present. We do this by defining the following effect

$$\begin{aligned}\text{eff}_{21}(\mathbf{w}) &= \frac{\text{Var}_{M_2}(\bar{y}_w)}{\text{Var}_{M_1}(\bar{y})}\\ &= \text{eff}_w(\mathbf{w})\,\text{eff}_I\left(1 + \rho_I\left[\bar{m}_I(\mathbf{w}) - 1\right] + \rho_C\left[\bar{m}_C(\mathbf{w}) - 1\right]\right),\end{aligned} \tag{2.15}$$

where

$$\text{eff}_I = \frac{n}{n + \rho_I\left(\sum_{i\in\mathcal{R}} n_i^2 - n\right)}.$$

The reference model (M_1) in $\text{eff}_{21}(\mathbf{w})$ models a simple random sample with an interviewer effect. Factor eff_I, indicates how close the corrected design effect to the survey effect is. For $\text{eff}_I = 1$ the corrected design and survey effect are equal and the closer eff_I is to zero the further apart are both effects. Hence, we can use eff_I to construct a measure for the contribution of the interviewer effect to the survey effect eff_{20}. For this, we first establish the following bounds for eff_I given in Result 2.

Result 2.

$$\frac{1}{n} \le \frac{n}{\rho_I(n - R)(n - R + 1) + n} \le \text{eff}_I \le \frac{R}{R + (n - R)\rho_I} \le 1.$$

The proof for Result 2 can be found in the Appendix.

Now we define a measure of the contribution of the interviewer effect to the survey effect inv_I as

$$\begin{aligned}\text{inv}_I\colon &\quad \left[\frac{1}{n}, 1\right] \mapsto [0, 1],\\ \text{inv}_I(a) &:= \frac{n(1 - a)}{n - 1} \quad \text{for} \quad \left[\frac{1}{n} \le a \le 1\right].\end{aligned} \tag{2.16}$$

For any given value of interviewer workloads $\mathbf{n}_I$, measure $\text{inv}_I(\text{eff}_I)$ is strictly increasing with decreasing eff_I. The maximum of $\text{inv}_I(\text{eff}_I)$ occurs at $R = 1$ and $\rho_I = 1$, which occurs when there is only one interviewer that always produces the same measurement. The minimum of $\text{inv}_I(\text{eff}_I)$ occurs at $\rho_I = 0$ for any given value of $\mathbf{n}_I$. If the concentration of the distribution of the workload over the interviewers increases and ρ_I stays fixed, $\text{inv}_I(\text{eff}_I)$ also increases. This relation becomes clearer if we write

$$\text{eff}_I = \frac{1}{1 + \rho_I\left(H_{\mathbf{n}_I} n - 1\right)}. \tag{2.17}$$

Alternatively, the coefficient of variation for the interviewer workloads $\text{CV}_{\mathbf{n}_I} = R\sigma_{\mathbf{n}_I}/n$, with $\sigma^2_{\mathbf{n}_I} = 1/R\sum_{i\in\mathcal{R}} n_i^2 - (n/R)^2$, could also be used to describe eff_I, since $H_{\mathbf{n}_I} = \left(1 + \text{CV}^2_{\mathbf{n}_I}\right)/R$ (Lynn and Gabler, 2004). Note that for $\sigma^2_{\mathbf{n}_I} = 0$ we have $\text{eff}_I = R/(R + (n - R)\rho_I)$.

Using Results 1 and 2, as well as inequality (2.12), we can give the following bounds for the corrected design effect.

Result 3.

$$\frac{n}{\rho_I\,(n-R)(n-R+1)+n}\,\mathrm{eff}_{20}^{*}\,(\mathbb{I}_n) \le \mathrm{eff}_{21}^{*}\,(\mathbf{w}) \le \mathrm{eff}_w\,(\mathbf{w})\,\frac{R}{R+(n-R)\,\rho_I}\,\mathrm{eff}_{20}^{*}\,(\mathbb{I}_n),$$

where eff_{21}^{*} is the corrected design effect when there are equal interviewer workloads and equal PSU sizes. The bounds of eff_{21}^{*}, given in Result 3 , do not depend on $\mathbf{n}_I$, but it should be noted that in the lower bound of $\mathrm{eff}_{21}^{*}\,(\mathbf{w}),\ \mathrm{eff}_I$ takes on its value for the maximum concentration in $\mathbf{n}_I$, whereas $\mathrm{eff}_{20}^{*}\,(\mathbb{I}_n)$ corresponds to the minimal concentration of $\mathbf{n}_I$. Since eff_I does not depend on $\mathbf{w}$, an upper (or lower) bound for $\mathrm{eff}_{21}\,(\mathbf{w})$ can be found by obtaining the upper (or lower) bounds of $\bar{m}_I\,(\mathbf{w})$, $\bar{m}_C\,(\mathbf{w})$ and $\mathrm{eff}_w\,(\mathbf{w})$ as described in the Appendix.

Finally, we introduce a corrected design effect that assumes the measurement model (M_{2*}), given by

$$\begin{aligned}\mathrm{eff}_{2*1}\,(\mathbf{w}) &= \frac{\mathrm{Var}_{M_{2*}}\,(\bar{y}_w)}{\mathrm{Var}_{M_1}\,(\bar{y})}\\ &= \mathrm{eff}_w\,(\mathbf{w})\,\mathrm{eff}_I\,(1+\rho_I\,[\bar{m}_I\,(\mathbf{w})-1]+\rho_C\,[\bar{m}_C\,(\mathbf{w})-1]+\rho_{IC}\,[\bar{m}_{IC}\,(\mathbf{w})-1]). \qquad (2.18)\end{aligned}$$

Similarly to Result 3 we can establish the following bounds for $\mathrm{eff}_{2*1}\,(\mathbf{w})$.

Result 4.

$$\frac{n}{\rho_I\,(n-R)(n-R+1)+n}\,\mathrm{eff}_{2*0}^{*}\,(\mathbb{I}_n) \le \mathrm{eff}_{2*1}^{*}\,(\mathbf{w}) \le \mathrm{eff}_w\,(\mathbf{w})\,\frac{R}{R+(n-R)\,\rho_I}\,\mathrm{eff}_{2*0}^{*}\,(\mathbb{I}_n).$$

Here eff_{2*1}^{*} corresponds to the case where n_{qi}, the number of respondents that belong to the q^{th} PSU and are interviewed by the i^{th} interviewer, is a constant, i.e., $n_{qi} = n/(RK)$ for all $i \in \mathcal{R}$ and $q \in \mathcal{K}$. This also implies that for eff_{2*1} we have $n_i = n/R$ and $n_q = n/K$. The proof of Result 4 can be found in the Appendix. Using model (M_{2*}) instead of (M_2) gives some additional flexibility in fitting the measurement model to the observed data. Whether this is required is a part of Section 3.2, where the different measurement models are tested against each other for ESS6 data.

3 Empirical findings from the ESS

After we established the effects associated with interviewers and multi-stage or cluster sampling, we now estimate the survey effect and our proposed corrected design effect for ESS6 data (ESS, 2016).

There were 29 participating countries in ESS6 (ESS, 2018a), but not all have been considered in our analysis. We excluded all countries with a single-stage design (there were no single-stage cluster sampling designs in ESS6). In addition, we excluded those countries that had a multi-domain sampling design.

These countries employed different sampling designs in different regions of the country, but they all refer to a certain level of the Nomenclature of Territorial Units for Statistics (NUTS), as established by Eurostat (ESS, 2013, pages 21-22). For example, Norway used a single stage sample for its more densely populated regions, which, combined, contained almost 75 percent of the target population, and a two-stage sampling design for the rest of the country.

First, in Section 3.1 we assess whether the estimation of the measurement models described in Section 2 is generally feasible, given the PSU-interviewer structure found in ESS6. To this end, we use a model-based simulation study. In Section 3.2, we test the different measurement models against each other in order to use the most appropriate ones for the estimation of the survey effect and the corrected design effect. Afterwards, we compare our results with the design effect that was used by the ESS to plan the sample size.

The PSU and interviewer identification variables needed for our simulation study and the estimation of the effects were obtained from the so-called Sampling Design Data Files (SDDFs) and the Interviewer Questionnaire, respectively (ESS, 2014a). The SDDFs contain information on the sampling design, including a PSU identifier. For ESS6, the SDDFs have to be downloaded individually for each country (ESS, 2018b).

3.1 Simulation for the stability assessment of effect estimates

Interviewers and sampling have long been recognized as principal sources of survey error. The way interviewers are deployed during fieldwork makes it difficult to separate the interviewer variance from the PSU variance. To make data collection more efficient, interviewers are usually assigned to work exclusively in certain regions (Von Sanden, 2004, Section 1.3). Correspondingly, interviewers in ESS6 seldom work across regions. For ESS6, we observe the following situation: In general, interviewers work in a number of PSUs within a certain area, but never in all PSUs. PSUs might be visited by more than one interviewer, but never by all of them. For 25% of all considered countries, the mean number of regions (variable *region*, ESS (2013), pages 21-22) an interviewer visited was 1.017 or lower. For 75% of all countries, the mean number of regions per interviewer was 1.256 or lower.

The non-hierarchical structure of PSUs and interviewers can be considered typical of large scale social surveys like the ESS. A so-called *fully interpenetrated* survey design, where all interviewers work in all PSUs, is in general unfeasible for country-wide surveys. This makes it difficult to decide what amount of observed similarity between observations made by an interviewer is due to intra-interviewer correlation or instead due to intra-PSU correlation. This problem has been addressed in a number of studies. For instance, by using a fully nested survey design, where multiple interviewers work in the same PSU but not across them (Schnell and Kreuter, 2005). But also so-called *partially interpenetrated* surveys, where different interviewers work in multiple PSUs and PSUs are visited by multiple interviewers, have been analyzed, (Davis and Scott, 1995; O’Muircheartaigh and Campanelli, 1998). These partially interpenetrated surveys resemble more the situation we observe for ESS6.

To test our measurement models and to disentangle the different variance components, we fit a multilevel model with crossed random effects. In another context, Raudenbush (1993) proposes to allow for so-called *crossed effects* in the random effects structure. These crossed effects allow for the situation of partially interpenetrated factors, and are able to estimate all three variance components of measurement model (M_{2*}), σ_I^2, σ_C^2, and σ_{IC}^2.

Vassallo, Durrant and Smith (2017) show, using simulations on synthetic data, how well a multilevel model with crossed random effects for cluster and interviewer can estimate the variance-covariance structure of the data model under different patterns of interpenetration between cluster and interviewer. They identify the sample size, the number of interviewers and PSUs, and the level of interpenetration as the driving factor for the quality of the estimates of the variance components. The level of interpenetration plays a decisive role for the quality of the variance component estimates. Vassallo et al. (2017) found that already 2-3 interviewers per PSU lead to relatively stable estimates of the variance components. However, their survey designs were all balanced and symmetric, meaning that the interpenetration of PSUs by interviewers was constant for all PSUs and vice versa. This is not the case for countries in ESS6. Therefore, we perform a simulation to test whether under the partial interpenetrated survey designs of ESS6 the variance components of our measurement model (M_2) can be estimated or not.

For the simulation, we generate samples from a n-dimensional multi-variate normal distribution $\text{MVN}(\boldsymbol{\mu}, \Sigma)$. The vector of means $\boldsymbol{\mu}$ contains, for each dimension, the same value. The covariance matrix Σ follows the variance-covariance structure of measurement model (M_2) and was constructed for each country based on the observed PSU-interviewer structure. The variance components were set to $\sigma_I^2 = 0.2$, $\sigma_C^2 = 0.08$, $\sigma^2 = 2$. We generated 1,000 samples from the superpopulation model $\text{MVN}(\boldsymbol{\mu}, \Sigma)$ for each country and estimated measurement model (M_2) for each of these samples. The simulation was implemented in *R* (R Core Team, 2019). The samples for the simulation were generated with the help of the *mvtnorm* package (Genz, Bretz, Miwa, Mi and Hothorn, 2019) and the estimation of the model was done using the *lme4* package (Bates, Mächler, Bolker and Walker, 2015, 2019).

Table 3.1 depicts the relative Monte Carlo bias of the estimators for the variance components of model (M_2). For an estimator $\hat{\theta}$ of θ we define this measure as

$$\text{MC-RBias } \hat{\theta} = \frac{\bar{\hat{\theta}}_{\text{MC}}}{\theta} - 1,$$

where $\bar{\hat{\theta}}_{\text{MC}} = \sum_{d=1}^{D} \hat{\theta}_d \big/ D$, θ is the true value, $\hat{\theta}_d$ the value of $\hat{\theta}$ for the d^{th} sample of the simulation and D is the total number of samples generated, i.e., $D = 1{,}000$, in our simulation. We see that σ_I^2 and σ_C^2 are estimated with a relative low bias for all considered countries in ESS6. In addition to the relative Monte Carlo bias, we have added the number of PSUs K, the number of interviewers R, the sample size n, the average number of PSUs that an interviewer works in $\bar{K}_I$, and the average number of interviewer that work in a PSU $\bar{R}_C$ to Table 3.1. $\bar{K}_I$ and $\bar{R}_C$ are used as measures for the level of interpenetration of PSUs by interviewers and interviewers by PSUs, respectively. For all countries, other than Germany, there are more PSUs than interviewers and $\bar{K}_I$ is greater than $\bar{R}_C$. $\bar{K}_I$ reaches from 1.423 in Germany to 17.396 in Albania. The level of $\bar{K}_I$ observed for all countries seems to be high enough to disentangle the variance components of model (M_2). We can observe a negative relationship between $\bar{K}_I$ and MC-RBias $\hat{\sigma}_I^2$, which can be mediated by n and K. Higher n and K correspond to a higher accuracy of $\hat{\sigma}_I^2$. An analogous observation can be made for MC-RBias $\hat{\sigma}_C^2$. A higher $\bar{R}_C$ also improves the precision of the estimates and can compensate for a low $\bar{K}_I$. A high enough one-sided interpenetration, either of the PSUs by the interviewers or vice versa, is sufficient to accurately estimate σ_I^2 and σ_C^2 for model (M_2). For example, the Czech Republic, which has the lowest $\bar{R}_C$, but a $\bar{K}_I$ of round 1.848, enables relative precise estimates for the variance components.

It should be noted that for measurement model (M_{2*}), both $\bar{K}_I$ and $\bar{R}_C$ are of importance. For example, σ_C^2 and σ_{IC}^2 cannot be estimated with precision if $\bar{R}_C$ is too low. For example, in a similar simulation for model (M_{2*}), it was not possible to obtain accurate estimates of σ_C^2 and σ_{IC}^2 for the Czech Republic, although the relative bias of $\hat{\sigma}_I^2$ was around 1 percent.

For Bulgaria and Czech Republic $\bar{R}_C = 1$, that is, their PSUs are nested within the interviewers. In this case, we do not have crossed random effects, but nested random effects, as we never have the case where respondents are within the same PSU but not interviewed by the same interviewer. For this special case, strictly speaking, σ_C^2 should be labeled σ_{IC}^2. But, for simplicity, for both cases we use σ_C^2 as a label for the variances of the PSU random effect. This is not entirely unjustified, as σ_{IC}^2 defines the additional correlation between respondents that are in the same PSU, compared to those respondents that are interviewed by the same interviewer, but are in different PSUs.

Table 3.1
Relative bias of random effect variance estimates

	MC-RBias $\hat{\sigma}_I^2$	MC-Bias $\hat{\sigma}_C^2$	K	R	n	$\bar{K}_I$	$\bar{R}_C$
Albania	0.00	-0.02	264	53	1,201	17.40	3.49
Belgium	0.00	-0.02	363	155	1,869	3.00	1.28
Bulgaria	-0.01	0.04	400	247	2,260	1.63	1.00
Czech Republic	0.01	-0.01	426	231	2,009	1.85	1.00
France	0.01	0.01	267	165	1,968	1.99	1.23
Germany	0.01	-0.00	156	194	2,958	1.42	1.77
Ireland	-0.01	0.01	212	116	2,628	2.15	1.17
Israel	-0.00	0.01	190	114	2,508	3.00	1.80
Italy	-0.02	0.05	129	117	960	1.49	1.35
Kosovo	0.01	-0.02	160	72	1,295	2.29	1.03
Slovakia	-0.02	0.04	249	132	1,847	1.93	1.02
Slovenia	-0.01	0.00	150	50	1,257	3.30	1.10
Spain	-0.01	0.03	422	74	1,889	8.20	1.44
Ukraine	0.00	0.00	306	237	2,178	1.44	1.11
United Kingdom	-0.01	0.00	226	150	2,286	2.36	1.57

Our simulation study confirms and extends the findings of Vassallo et al. (2017) for the unbalanced situation of the ESS6. We also saw that the PSU-interviewer structure observed for ESS6 does not prohibit the disentanglement of σ_C^2 and σ_I^2 for measurement model (M_2).

3.2 Survey effects in ESS round 6

As seen in our simulation study, the estimation of the interviewer and cluster variance is feasible in ESS6. Now we test, for a set of selected variables from the ESS main questionnaire (ESS, 2013), each variance component of model (M_{2*}) on its significance. All used variables, except age and gender, have an ordinal scale, but are treated as metric variables for the purpose of this analysis. A list of all used variables can be found in the Appendix.

As a variance component has its minimum at zero, the test is performed on the boundary of the parameter space, which imposes classical problems from test theory. Scheipl, Greven, and Kuechenhoff (2008) proposed a restricted likelihood ratio test, designed to test for a zero random effects variance. We use their implementation of this test in the R-Package *RLRsim* and perform three test decisions.

First, we test on the significance of the interaction variance of interviewers and PSUs, when assuming relevant interviewer and PSU variances. Our null hypothesis is H_0: $\sigma_{IC}^2 = 0$ versus alternative hypothesis H_A: $\sigma_{IC}^2 > 0$. The per country average of rejected null hypothesis over the different variables is displayed in Table 3.2. The first two columns correspond to two different type I error levels for the test of H_0: $\sigma_{IC}^2 = 0$, indicated by $\alpha = 0.01$ and 0.05. Israel is the country that has the highest number for significant interaction variance σ_{IC}^2 on all type I error levels. For all other countries the null hypothesis is not rejected for all variables at a significance level of 1%. Although not displayed in Table 3.2 it can be noted that at a 10% significance level two-thirds of the countries have at least some variables with a significant interaction variance. Therefore, the possibility of an interaction effect should be considered when estimating survey effects.

In our second test decision an interviewer variance but no interaction variance is assumed. The null hypothesis is that the PSU variance is not relevant, that is H_0: $\sigma_C^2 = 0$ versus the alternative hypothesis H_A: $\sigma_C^2 > 0$. Average test results for the different type I error levels can be found in the columns 3 to 4 of Table 3.2. For some variables, the PSU variances are not significant as an addition to the interviewer variance. This result is especially strong for Belgium, where only 3% of the variables seem to have a PSU variance. However, also for France and Slovenia, the PSU variance is only significant at a level of 1% for a relative small number of the variables and for Albania for none of the variables. In contrast to that, Bulgaria, Ireland, Israel and Slovakia have significant PSU variance for the majority of variables. Overall, the PSU variance appears to be relevant in most countries and thus should be considered when estimating survey effects.

For the third test decision we perform, a PSU variance but no interaction variance is assumed. The null hypothesis is that the interviewer effect is not relevant H_0: $\sigma_I^2 = 0$ versus the alternative hypothesis H_A: $\sigma_I^2 > 0$. Average test results can be found in columns 5 to 6 of Table 3.2. The lowest rejection rates are found in Germany and France, although 19% of the variables for Germany and 23% for France still have a significant interviewer variance at a 1% significance level. The other countries show a far higher proportion of variables with significant interviewer variance. On the 1% and 5% significance level, the interviewer variance has a higher rejection rate than the PSU variance for 13 out of the 15 countries. Thus, the interviewer variance appears to be of relevance for all countries in ESS6, indicating that possible interviewer effects should be taken into account when assessing the efficiency of survey designs.

Table 3.2
Rejection rates for existence of variance components

H_0:	$\sigma_{CI}^2 = 0$		$\sigma_C^2 = 0$		$\sigma_I^2 = 0$	
α	**0.01**	**0.05**	**0.01**	**0.05**	**0.01**	**0.05**
Albania	0.00	0.03	0.00	0.16	0.55	0.77
Belgium	0.00	0.00	0.03	0.03	0.77	0.90
Bulgaria	0.00	0.00	0.81	0.90	0.90	1.00
Czech Republic	0.00	0.00	0.52	0.58	1.00	1.00
France	0.00	0.00	0.10	0.23	0.23	0.45
Germany	0.00	0.00	0.26	0.61	0.19	0.42
Ireland	0.00	0.06	0.77	0.81	0.94	0.97
Israel	0.13	0.32	0.94	1.00	0.84	0.94
Italy	0.00	0.03	0.10	0.32	0.42	0.65
Kosovo	0.00	0.00	0.45	0.58	0.94	0.97
Slovakia	0.00	0.00	0.77	0.90	0.97	0.97
Slovenia	0.00	0.00	0.03	0.16	0.74	0.84
Spain	0.00	0.00	0.13	0.23	0.74	0.84
Ukraine	0.00	0.00	0.55	0.74	0.90	0.94
United Kingdom	0.00	0.03	0.19	0.35	0.71	0.87

Based on the selected models for the different variables, survey effects defined in equation (2.8) are estimated. Table 3.3 shows the country specific average of estimated survey effects over all considered variables. In addition Table 3.3 also contains the average of design effect deff, as it is used by the ESS to plan sample sizes. In our notation this design effect has the form

$$\text{deff} = \text{eff}_w \left(1 + \rho_C \left(\overline{m}_C(\mathbf{w}) - 1\right)\right).$$

To estimate ρ_C in deff we used an ANOVA estimator (The ESS Sampling Expert Panel, 2016; Ganninger, 2010, page 45) and do not test for the significance of the PSU variance. Measurement model a used in eff_{a0} can include interviewer, PSU and interaction variance, if the model selection identifies it as significant at a level of 0.05. The same applies to measurement model a used in eff_{a1}, i.e., the corrected design effect. If interviewer variance is identified as not significant for a variable, then eff_{a1} becomes eff_{a0}. To measure the influence of the interviewer on the survey effect inv_I is also shown.

By comparing deff and eff_{a1} in Table 3.3 an interesting observation can be made: For Germany deff is clearly lower than for Ireland and the Czech Republic. From this we could deduce that Germany would need a much lower sample size to achieve the same average effective sample size as Ireland and the Czech Republic. However, if we look at eff_{a1}, this relation switches. Table 3.3 shows that the cluster effect of the complex sampling design is higher in Germany than it is in Ireland or the Czech Republic. Meaning that, if we are interested in equal average effective sample across countries, Germany would need a higher sample size than in Ireland or the Czech Republic. For example, for the Czech Republic to achieve an effective sample size of 1,500 with the standard design effect deff from Table 3.3 we would plan with a net sample of round 3,925 and for Germany with one of 3,115. If instead we use the corrected design effect eff_{a1}, to base the planning of the net sample size solely on the effect of the sampling design, we would select a net sample size of round 1,707 and 2,598, for the Czech Republic and Germany, respectively. This finding is also reflected in the values of inv_I, which indicates that a large part of eff_{a0} for Ireland and the Czech Republic can be attributed to an interviewer effect, whereas for Germany, the interviewer effect is smaller and eff_{a0} seems to be dominated by the cluster effect. Apart from Israel, Slovakia, and Slovenia, all countries have different ranks for deff and eff_{a1}, indicating that the allocation of the sample size over all countries would be very different, if the corrected design was used to plan effective samples sizes, instead of the conventional design effect deff.

Table 3.3
Average effect sizes for ESS6

	deff	$\textbf{eff}_{a0}$	$\textbf{eff}_{a1}$	$\textbf{inv}_I$
Albania	2.07	2.87	1.68	0.35
Belgium	1.18	1.75	1.01	0.37
Bulgaria	2.32	3.88	1.21	0.65
Czech Republic	2.62	6.58	1.14	0.78
France	1.69	1.80	1.46	0.16
Germany	2.08	2.28	1.73	0.19
Ireland	3.32	5.42	1.26	0.73
Israel	2.41	4.67	1.42	0.61
Italy	1.76	2.20	1.32	0.34
Kosovo	4.01	10.97	1.51	0.80
Slovakia	5.02	20.28	2.27	0.85
Slovenia	1.59	3.03	1.06	0.55
Spain	1.16	2.01	1.05	0.42
Ukraine	2.97	5.61	1.18	0.73
United Kingdom	1.76	2.24	1.32	0.38

deff: average design effects as defined in equation (1.1).
eff_{a0}: average survey effect with measurement model of interest (M_a) and (M_0) as reference.
eff_{a1}: average corrected design effects with measurement model of interest (M_a) and (M_1) as reference.
inv_I: average contribution of interviewer effects to the design effect as defined in equation (2.16).

eff_{a1} is smaller than deff for all countries, and their distance, $|\text{deff} - \text{eff}_{a1}|$, has a positive but non-linear relationship with inv_I. The lowest values of $|\text{deff} - \text{eff}_{a1}|$ are observed for Spain, Belgium, France, and Germany, which are all countries whose inv_I value is below the median of inv_I. The opposite is observed for Slovakia, Kosovo, Ireland, and Ukraine, the countries with the highest distance between deff and eff_{a1}. These countries all have a value of inv_I that is higher than the median value of inv_I. These patterns for countries with a relatively high distance between deff and eff_{a1} are consistent with what we would expect if there is a high interviewer effect present in the data. The opposite can be said for countries when a relatively small distance between deff and eff_{a1} is observed.

Interviewer effects depend on many different factors (West and Blom, 2017), including the type of the question asked and the used ESS6 data is mostly gathered from attitude questions. Hence, the presented results in this section cannot be extrapolated to other types of surveys in the same countries.

4 Conclusions

Using a design effect to select a sample size is a commonly used method to account for the loss of efficiency that a complex sampling design might entail. However, the design effect can be inflated by an interviewer effect in face-to-face surveys. This can lead to erroneous conclusions about the effect that complex sampling has on the efficiency of a sampling strategy. As a consequence, this could lead to misallocation of resources. The planned sample size might be too high, if it is based on an overestimated design effect. Therefore, we propose to consider both the design and the interviewer effect simultaneously when planning a sample size. The survey effect, which we develop in Section 2, accounts both for interviewer and PSU variance to assess the efficiency of a survey design. Based on the survey effect we introduce a corrected design effect, which uses as a reference design a simple random sample with an interviewer effect. As a result, the corrected design effect is no longer conflated with the interviewer effect and can be used to better base the decision on the samples size on the effect the sampling design has on the precision of survey estimates.

For ESS6, our empirical findings in Section 3.2 show that high design effects are related to high interviewer effects. The average corrected design effects that we observe suggest that the sampling design influences the variance of an estimator to a lesser degree than interviewers for many countries in the ESS6. The ability to estimate the corrected design effect, e.g., from historical data as guide for the survey planner, depends mainly on the PSU-interviewer structure and the allocation of interviewer workloads and cluster sizes. We find a partially interpenetrated survey design, i.e., on a regional level, can be sufficient to disentangle PSU and interviewer variance. In our simulation study an average number of 1.5 PSUs per interviewer or interviewers per PSU was enough to estimate the variance components of measurement model (M_2). For actual survey data, that is categorical, this level of interpenetration might not be high enough, but a high number of PSUs, interviewers, and a large sample size might off-set a low interpenetration. For practical applications, we recommend testing via simulation if the assumed measurement model can be estimated with the given PSU-interviewer structure, as we did in Section 3.1.

When using the survey effect and corrected design effect for the planning of a sample size it can be helpful to work with the upper and lower bounds of these statistics. In Section 2, we derive such bounds, but under somewhat unrealistic assumptions regarding the distribution of survey weights, interviewer workloads and PSU sizes. However, if realistic assumptions about the concentration of survey weights, interviewer workloads and PSU sizes can be made, then we propose to use a linear optimization, as shown in the Appendix, to derive bounds that are of much higher practical relevance and can serve as valuable guidance for survey planners. Generally, we recommend to have lowly concentrated distributions of interviewer workloads and PSU cluster sizes in order to increase the precision of survey estimates. Thus, interviewer workloads and PSU cluster sizes should be as equal as possible for any given number of interviewer and PSUs.

The measurement models we introduce in Section 2 are arguably simplistic. This makes the models applicable to most survey designs. The only information, besides the survey data, used to compute the estimates for Table 3.3 were the PSU and interviewer indicators. However, there are certain aspects of survey measurements that could be incorporated into a practical measurement model, such as stratification, which, in general, increases the efficiency of an estimation strategy (Särndal et al., 1992, Section 3.7). This was neglected in our analysis, despite the fact that many ESS6 countries used a stratified design for their PSU sample. Gabler, Häder and Lynn (2006) develop a design effect for estimation strategies that combine different sampling designs for sampling domains. This approach could possibly be adapted to add a stratification effect to the PSU variance. Furthermore, it might be plausible to assume that interviewers differ with regard to the degree of homogeneity that they add to their measurements. This interviewer heterogeneity could be incorporated into a measurement model by allowing groups of interviewers to have different distributions of $\mathfrak{I}_i$, i.e., values for σ_I^2 (West and Elliott, 2014). However, a procedure to classify interviewers would be needed. Preferably one that does mainly rely on the survey data and not so much on information available about the interviewers, which might differ from survey to survey.

A future application for the presented framework of the survey effect would be to find an optimal budget allocation with respect to the number of PSUs and interviewers, for a given effective sample size. Such an optimization requires a cost model for the deployment of interviewers to a possible set of PSUs. Fieldwork institutes could possibly provide the necessary information to calculate such a model for a particular country. Such a method could help survey planners to conduct face-to-face surveys more effectively, which is of increasing importance as surveys based on probability samples are under pressure from the comparably cheap alternative of recruiting respondents from online-access panels.

Further research could also focus on the development of survey effect for other estimators than the weighted sample mean. For estimators that can be described as functions of estimated totals, which includes the Ordinary Least Square Estimator for regression coefficients (Särndal et al., 1992, Section 5.10), it should be possible to derive survey effects, under the framework shown in Section 2, that allow for a similar factorization as the survey effect presented in this work.

Appendix

For the Appendix we will introduce a short notation of multiple sums, where, for example, $\sum_{qik} y_{qik}$ will be shorthand for

$$\sum_{q \in \mathcal{K}} \sum_{i \in \mathcal{R}} \sum_{k \in s_{qi}} y_{qik}.$$

Results 1

$$\text{eff}^*_{20}(\mathbf{w}) \leq \frac{n \sum_{qik} w^2_{qik}}{\left(\sum_{qik} w_{qik}\right)^2} \left(1 + \rho_I \left[\frac{n}{R} - 1\right] + \rho_C \left[\frac{n}{K} - 1\right]\right).$$

Proof: We need to show that

$$\frac{\sum_i \left(\sum_{qk} w_{qik}\right)^2}{\sum_{qik} w^2_{qik}} \leq \frac{n}{R} \tag{A.1}$$

and

$$\frac{\sum_i \left(\sum_{qk} w_{qik}\right)^2}{\sum_{qik} w^2_{qik}} \leq \frac{n}{K} \tag{A.2}$$

hold, if $n_i = \frac{n}{R}$ and $n_q = \frac{n}{K}$, for all $i = 1, \ldots, R$ and $q = 1, \ldots, K$.

As shown in Gabler et al. (1999), if $a_{qik} = 1$ for all $q \in \mathcal{K},\ i \in \mathcal{R},\ k \in s_{qi}$, using the Cauchy-Schwarz inequality, we know that

$$\left(\sum_{qk} w_{qik} a_{qik}\right)^2 = \left(\sum_{qk} w_{qik}\right)^2 \leq n_i \sum_{qk} w^2_{qik} = \sum_{qk} a^2_{qik} \sum_{qk} w^2_{qik}$$

$$\frac{\sum_i \left(\sum_{qk} w_{qik}\right)^2}{\sum_i \sum_{qk} w^2_{qik}} \leq \frac{\sum_i n_i \sum_{qk} w^2_{qik}}{\sum_i \sum_{qk} w^2_{qik}}.$$

If we have $n_i = \frac{n}{R}$ for all $i = 1, \ldots, R,$ then it follows that

$$\frac{\sum_i \left(\sum_{qk} w_{qik}\right)^2}{\sum_{qik} w^2_{qik}} \leq \frac{n}{R}.$$

The proof for inequality (A.2) is analogous to the one above, which completes the proof of Result 1.

Upper bounds for $\bar{m}_I(\mathbf{w}), \bar{m}_C(\mathbf{w})$ and $\text{eff}_w(\mathbf{w})$

For given $\mathbf{n}_I^\top$ and $\mathbf{n}_C^\top$ and $w_k \in [a, b]$ with $a, b \in \mathbb{R}_+$ for all $k \in s,$ and $\sum_k w_k = n$ we can construct an upper bound for $\bar{m}_I(\mathbf{w})$ and $\bar{m}_C(\mathbf{w})$.

We know that

$$\frac{\sum_i \left(\sum_{qk} w_{qik}\right)^2}{\sum_i \sum_{qk} w^2_{qik}} \leq \frac{\sum_i n_i \sum_{qk} w^2_{qik}}{\sum_i \sum_{qk} w^2_{qik}} \leq \frac{\sum_i n_i \sum_{qk} w^2_{qik}}{n}. \tag{A.3}$$

Now we need to find a sufficiently high value for $\sum_i n_i \sum_{qk} w_{qik}^2$. For this we define $x_i = \sum_{qk} w_{qik}^2$ and $\mathbf{x} = (x_1, \ldots, x_I)^\top$. Thus we have to solve the following problem:

$$\begin{aligned}
&\max_{\mathbf{x}\in \mathrm{R}^R} \mathbf{n}_I^\top \mathbf{x} \\
&\text{s.t.} \\
&x_i \geq a^2 n_i \ \ \forall i \in \mathcal{R} \\
&x_i \leq b^2 n_i \ \ \forall i \in \mathcal{R} \\
&\sum_i x_i \geq n \\
&\sum_i x_i \leq f_{sqm}(a, b, n),
\end{aligned} \tag{A.4}$$

where

$$f_{sqm}(a, b, n) = b^2 \left\lfloor \frac{n - nb}{a - b} \right\rfloor + (n - nb) - \left\lfloor \frac{n - nb}{a - b} \right\rfloor (a - b) + b + a^2 \left(n - \left\lfloor \frac{n - nb}{a - b} \right\rfloor - 1 \right),$$

where $\lfloor \ \rfloor$ means rounded to the nearest lower integer. The problem formulated in equation (A.4) can be solved using a solver for linear programs, e.g., with the *solveLP* function from the *R* package Henningsen (2012). Function f_{sqm} gives a maximum of $\sum_k w_k^2$ given the upper and lower bounds of the weights a and b and the fact that the weights are scaled to n, i.e., $\sum_k w_k = n$. The sum of squares is maximized by giving as many weights their highest possible value b under the condition that each weight must have at least a value of a and that $\sum_k w_k = n$. The problem can then be solved using a simplex algorithm. An upper bound for $\overline{m}_C$ can be determined in the same fashion. Changing the problem to minimization and a lower bound for eff_{20} can be found. However, it is not guaranteed that separate optimization of $\overline{m}_C$ and $\overline{m}_I$ will yield values of $\mathbf{x}$ that allow for a value of $\mathbf{w}$ that jointly maximizes (or minimizes) $\overline{m}_C$ and $\overline{m}_I$. Although, if, $\mathbf{x}_C$ and $\mathbf{x}_I$ are the vectors that optimizes $\overline{m}_C$ and $\overline{m}_I$ respectively, it should be possible to find a possible value for $\mathbf{w}$, e.g., using iterative proportional fitting.

For $\text{eff}_w(\mathbf{w})$ we have under the same assumptions as made above

$$1 \leq \text{eff}_w(\mathbf{w}) = \frac{\sum_{k\in s} w_k^2}{n} \leq \frac{f_{sqx}(a, b, n)}{n}. \tag{A.5}$$

Result 2

$$\frac{n}{\rho_I (n - R)(n - R + 1) + n} \leq \text{eff}_I \leq \frac{R}{R + (n - R)\rho_I}.$$

Proof: The upper bound in Result 2 can be shown by using the Cauchy-Schwarz inequality, which gives us

$$\begin{aligned}
R \sum_i n_i^2 &\geq \left(\sum_i n_i \right)^2 \\
\sum_i n_i^2 &\geq \frac{n^2}{R}.
\end{aligned} \tag{A.6}$$

With a some algebra we can formulate the upper bound of eff_I.

To prove the lower bound in Result 2 we solve the following problem:

$$\begin{aligned}&\max_{\mathbf{n}_I \in \mathrm{N}^R_{>0}} \mathbf{n}_I^{\mathrm{T}} \mathbf{n}_I \\ &\text{s.t.} \\ &\sum_i n_i = n.\end{aligned} \tag{A.7}$$

A solution to the problem formulated in (A.7) can be found by considering that if we have $n_i - 1 \geq 1$ and $n_i \leq n_j$ it follows that $(n_i - 1)^2 + (n_j + 1)^2 > n_i^2 + n_j^2$. Thus for $n_j = \max_{i \in \mathcal{R}} n_i$ we can increase $\sum_i^R n_i^2$ if we reduce any $n_i > 1$ $i \neq j$ by one and add one to n_j. Hence, if $n_i = 1$ for all $i \neq j \in \mathcal{R}$ and $n_j = n - R + 1$ then $\sum_i n_i^2$ is at its maximum, with $\sum_i n_i^2 = (R - 1) + (n - R + 1)^2$.

Result 4

Proof: Given Result 2, to prove the right-hand side of Result 4 we need to show that

$$\mathrm{eff}^*_{2*0}(\mathbf{w}) \leq \frac{n \sum_{qik} w_{qik}^2}{\left(\sum_{qik} w_{qik}\right)^2} \left(1 + \rho_I \left[\frac{n}{R} - 1\right] + \rho_C \left[\frac{n}{K} - 1\right] + \rho_{IC} \left[\frac{n}{RK} - 1\right]\right). \tag{A.8}$$

To prove inequality (A.8) we only need to show that

$$\frac{\sum_{qi} \left(\sum_k w_{qik}\right)^2}{\sum_{qik} w_{qik}^2} \leq \frac{n}{RK}.$$

The rest follows from the proofs of inequalities (A.1) and (A.2). Thus it is sufficient to show that

$$\left(\sum_k w_{qik}\right)^2 = \left(\sum_k w_{qik} a_{qik}\right)^2 \leq n_{qi} \sum_k w_{qik}^2 = \sum_k a_{qik}^2 \sum_k w_{qik}^2,$$

if $a_{qik} = 1$ for all $q \in \mathcal{K}$, $i \in \mathcal{R}$, $k \in s_{qi}$, which also follows from the Cauchy-Schwarz inequality. Inequality (A.8) then follows if $n_{qi} = \frac{n}{RK}$ for $i = 1, \ldots, R$ and $q = 1, \ldots, K$.

The left-hand side of Result 4 follows from the proof of Result 6 in Gabler and Lahiri (2009) and Result 2.

ESS6 variables used for empirical evaluation

Table A.1
ESS6 variables used for empirical evaluation

pplfair	trstprt	stfdem	imueclt	iorgact
pplhlp	trstep	stfedu	imwbcnt	agea
polintr	trstun	stfhlth	happy	gndr
trstprl	lrscale	gincdif	aesfdrk	
trstlgl	stflife	freehms	health	
trstplc	stfeco	euftf	rlgdgr	
trstplt	stfgov	imbgeco	wkdcorga	

The definition of these variables including question text can be found in ESS (2013).

References

Bates, D.M., Mächler, M., Bolker, B.M. and Walker, S.C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67, 1, 1-48. https://doi.org/10.18637/jss.v067.i01.

Bates, D.M., Mächler, M., Bolker, B.M. and Walker, S.C. (2019). *Lme4: Linear Mixed-Effects Models Using 'Eigen' and S4*. https://CRAN.R-project.org/package=lme4.

Beullens, K., and Loosveldt, G. (2016). Interviewer effects in the European social survey. *Survey Research Methods*, 10, 2, 103-118.

Biemer, P.P. (2010). Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly*, Oxford University Press, 74, 5, 817-848.

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

Chaudhuri, A., and Stenger, H. (2005). *Survey Sampling: Theory and Methods*. CRC Press.

Davis, P., and Scott, A.(1995). The effect of interviewer variance on domain comparisons. *Survey Methodology*, 21, 2, 99-106. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/1995002/article/14405-eng.pdf.

Ellis, P.D. (2010). *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results*. Cambridge University Press.

European Social Survey (ESS) (2013). *ESS6 Data Protocol. 1.4*. London: ESS ERIC. http://www.europeansocialsurvey.org/data/download.html?r=6.

European Social Survey (ESS) (2014a). *European Social Survey Round 6 Interviewer Questionnaire*. Dataset edition: 2.1. London: ESS ERIC.

European Social Survey (ESS) (2014b). *Weighting European Social Survey Data*. London: ESS ERIC. www.europeansocialsurvey.org/docs/methodology/ESS_weighting_data_1.pdf.

European Social Survey (ESS) (2014c). *ESS6 – 2012 Documentation Report*. Edition: 2.3. London: ESS ERIC. http://www.europeansocialsurvey.org/docs/round6/survey/ESS6_data_documentation_report_e02_3.pdf.

European Social Survey (ESS) (2016). *European Social Survey Round 6 Data*. Dataset edition: 2.2. London: ESS ERIC.

European Social Survey (ESS) (2018a). *Countries by Round (Year)*. London: ESS ERIC. http://www.europeansocialsurvey.org/data/country_index.html.

European Social Survey (ESS) (2018b). *Data and Documentation by Round European Social Survey (ESS)*. London: ESS ERIC. http://www.europeansocialsurvey.org/data/download.html?r=6.

Fahrmeir, L., Heumann, C., Künstler, R., Pigeot, I. and Tutz, G. (1997). *Statistik: Der Weg Zur Datenanalyse*. 1sted. Berlin: Springer-Verlag.

Fischer, M., West, B.T., Elliott, M.R. and Kreuter, F. (2018). The impact of interviewer effects on regression coefficients. *Journal of Survey Statistics and Methodology*, May. https://doi.org/10.1093/jssam/smy007.

Gabler, S., Häder, S. and Lahiri, P. (1999). A model based justification of Kish's formula for design effects for weighting and clustering. *Survey Methodology*, 25, 1, 105-106. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/1999001/article/4718-eng.pdf.

Gabler, S., Häder, S. and Lynn, P. (2006). Design effects for multiple design samples. *Survey Methodology*, 32, 1, 115-120. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2006001/article/9256-eng.pdf.

Gabler, S., and Lahiri, P. (2009). On the definition and interpretation of interviewer variability for a complex sampling design. *Survey Methodology*, 35, 1, 85-99. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2009001/article/10886-eng.pdf.

Ganninger, M. (2010). *Design Effects: Model-Based Versus Design-Based Approach*. Edited by GESIS - Leibniz-Institut für Sozialwissenschaften. Array 3.

Genz, A., Bretz, F., Miwa, T., Mi, X. and Hothorn, T. (2019). *Mvtnorm: Multivariate Normal and T Distributions*. https://CRAN.R-project.org/package=mvtnorm.

Groves, R.M. (2009). *Survey Methodology*. 2nd ed. Wiley Series in Survey Methodology. Hoboken, New York: John Wiley& Sons, Inc.

Groves, R.M., and Lyberg, L. (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly*, Oxford University Press, 74, 5, 849-879.

Henningsen, A. (2012). *Linprog: Linear Programming/Optimization*. https://CRAN.R-project.org/package=linprog.

Kish, L. (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Association*, 57, 297, 92-115. https://doi.org/10.1080/01621459.1962.10482153.

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

Lohr, S.L. (2014). Design effects for a regression slope in a cluster sample. *Journal of Survey Statistics and Methodology*, 2, 2, 97-125. https://doi.org/10.1093/jssam/smu003.

Lynn, P., and Gabler, S. (2004). *Approximations to B* in the Prediction of Design Effects Due to Clustering*. ISER Working Paper Series.

Lynn, P., Häder, S., Gabler, S. and Laaksonen, S. (2007). Methods for achieving equivalence of samples in cross-national surveys: The European social survey experience. *Journal of Official Statistics*, 23, 1, 107.

O'Muircheartaigh, C., and Campanelli, P. (1998). The relative impact of interviewer effects and sample design effects on survey precision. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 161, 1, 63-77.

Raudenbush, S.W. (1993). A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. *Journal of Educational Statistics*, 18, 4, 321-349. https://doi.org/10.2307/1165158.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

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

Scheipl, F., Greven, S. and Kuechenhoff, H. (2008). Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. *Computational Statistics & Data Analysis*, 52, 7, 3283-3299.

Schnell, R., and Kreuter, F. (2005). Separating interviewer and sampling-point effects. *Journal of Official Statistics*, 21, 3, 389-410.

The ESS Sampling Expert Panel (2016). *Sampling Guidelines: Principles and Implementation for the European Social Survey*. London: ESS ERIC Headquarters. http://www.europeansocialsurvey.org/docs/round8/methods/ESS8_sampling_guidelines.pdf.

Vassallo, R., Durrant, G. and Smith, P. (2017). Separating interviewer and area effects by using a cross-classified multilevel logistic model: Simulation findings and implications for survey designs. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180, 2, 531-550.

Von Sanden, N.D. (2004). *Interviewer Effects in Household Surveys: Estimation and Design*. Ph.d. Thesis, Wollongong: University of Wollongong. http://ro.uow.edu.au/theses/312.

West, B.T., and Blom, A.G. (2017). Explaining interviewer effects: A research synthesis. *Journal of Survey Statistics and Methodology*, 5, 2, 175-211. https://doi.org/10.1093/jssam/smw024.

West, B.T., and Elliott, M.R. (2014). Frequentist and Bayesian approaches for comparing interviewer variance components in two groups of survey interviewers. *Survey Methodology*, 40, 2, 163-188. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2014002/article/14092-eng.pdf.

Vol. 46, No. 1, pp. 121-139
Statistics Canada, Catalogue No. 12-001-X

A new double hot-deck imputation method for missing values under boundary conditions

Yousung Park and Tae Yeon Kwon[1]

Abstract

In surveys, logical boundaries among variables or among waves of surveys make imputation of missing values complicated. We propose a new regression-based multiple imputation method to deal with survey nonresponses with two-sided logical boundaries. This imputation method automatically satisfies the boundary conditions without an additional acceptance/rejection procedure and utilizes the boundary information to derive an imputed value and to determine the suitability of the imputed value. Simulation results show that our new imputation method outperforms the existing imputation methods for both mean and quantile estimations regardless of missing rates, error distributions, and missing-mechanisms. We apply our method to impute the self-reported variable "years of smoking" in successive health screenings of Koreans.

Key Words: Hot-deck; Two-sided boundary conditions; Multiple imputation; Item nonresponse.

1 Introduction

Survey nonresponse (or item nonresponse) arises in many censuses or sample surveys, and several methods for filling in such missed items have been proposed. Some missing variables in a survey are logically bounded. For example, in the U.S. National Health Interview Survey, some families did not report exact income but did report income categories, which provides bounds of the exact family income values. When personal earnings within a family are reported for some family members but not for the others, the sum of the reported personal earnings gives a lower bound of the family income (Schenker, Raghunathan, Chiu, Makuc, Zhang and Cohen, 2006). Geraci and McLain (2018) addressed several examples of bounded missing variables in surveys which include psychometric scales, clinical scores, and school grades.

Waves in panel surveys and panel data sets often provide logical constraints for missing variables. In the periodic health screening data released by the national health service of Korea, the missing smoking period of a smoker at the current wave is bounded below by his/her smoking period reported at the previous wave and bounded above by his/her age. We observed from 2011 and 2013 health screening data of Korea that the 2013 data contain up to 73.5% missing values for smoking periods when we treat the smoking periods that violate such logical constraints as missing. In particular, the mean age of respondents to the smoking periods question is 9 years younger than that of non-respondents, implying that the missing mechanism of the smoking period is not missing completely at random (MCAR) and hence imputation is required.

Geraci and McLain (2018) proposed a quantile-based imputation method for one-sided or two-sided missing variables in which upper and lower values are fixed constants, and showed that their method had advantages, especially when the sample size is moderately large and the true model is strictly non-linear.

[1] Yousung Park, Professor, Department of Statistics, Korea University, 145 Anam-ro, Anam-dong, Seongbuk-gu, Seoul, Republic of Korea. E-mail: yspark@korea.ac.kr; Tae Yeon Kwon, Assistant Professor, Department of International Finance, Hankuk University of Foreign Studies, 81 Oedae-ro, Wangsan-ri, Mohyeon-myeon, Cheoin-gu, Yongin-si, Gyeonggi-do, Republic of Korea. E-mail: tykwon@hufs.ac.kr.

The most common way to accomodate the logical boundaries to a multiple imputation method is by adopting truncation or an acceptance/rejection step. However, our simulation studies shows that this additional step to existing multiple imputation methods introduces bias as long as the two-sided boundaries are asymmetric.

We propose a new regression-based multiple imputation method without an acceptance/rejection or truncation procedure. This new method utilizes the two-sided boundaries for imputing missing values, automatically meets the boundary constraints, and includes the imputation method given in Kwon and Park (2015) as a special case. We call this new method double hot-deck boundary information matching proportioned residual draw method (DBM-PRD), because two hot-deck steps are used to reduce the number of donor candidates and to choose an appropriate proportional residual that is defined by the usual residual divided by the distance between an observation and its lower or upper bound. This proportioned residual was used in Kwon and Park (2015).

Hot-deck imputation that replaces a missing value with a "similar" observation can improve the imputation performance relative to imputation methods that are derived only from model-assisted schemes. Andridge and Little (2010) showed that, in particular, when a model is used to define matches, hot-deck is less vulnerable to model misspecification than model-assisted methods.

Multiple imputation incorporates imputation uncertainty into statistical inference by substituting missed values several times. The basic method given in Rubin (1978) is to impute the missing value with a sampled value from the normal posterior distribution. It has been extended for imputation of missing values with a logical boundary (Raghunathan, Lepkowski, Van Hoewyk and Solenberger, 2001) by using a truncated normal posterior distribution (T-NORM). Rubin and Schenker (1986) and Rubin (2004) adopted the empirical distribution of the observed standardized residuals based on a fitted regression model. They proposed an imputation method adjusted for uncertainty of the mean and variance (MV), which imputes the missing value with its predictive mean plus residuals that are randomly chosen from their empirical distribution.

Extended from these basic ideas of hot-deck and multiple imputation, most existing methods assume the distribution of the data (usually normal) and employ a truncated distribution (usually truncated normal) to meet a logical boundary of the missing value (van Buuren and Groothuis-Oudshoorn, 2010; Honaker, King and Blackwell, 2012; Su, Gelman, Hill, Yajima, 2011; Raghunathan et al., 2001; Raghunathan, Solenberger and Van Hoewyk, 2002). The predictive mean matching method (PMM) imputes a missing value with a randomly selected observation having a similar predictive mean to that of missing value (Little, 1988). Schenker and Taylor (1996) proposed the local residual draw method (LRD), which replaces each missing value with its predictive mean plus a randomly drawn residual whose predictive mean is close to that of missing value. Instead of the residual in LRD, Kwon and Park (2015) used the proportioned residual whose distance from the predicted mean to its boundary value is close to that of the missing value in order to meet an one-sided boundary imposed on the variables of interest.

The DBM-PRD method is essentially along the same lines as Kwon and Park (2015). However, DBM-PRD employs one more matching procedure to take into account two-sided boundaries and to resolve asymmetric boundary information. This additional matching is based on which boundary is closer to the predictive mean of each missing value. Meanwhile, DBM-PRD imputes the missing value with its predictive mean plus a proportioned residual multiplied by the distance between the predicted mean and the corresponding upper or lower bound. Although DBM-PRD belongs to a mixed method as it uses a regression model in the first step and a double hot-deck imputation of missing values in the second step, the DBM-PRD method is new in that it directly adjusts for the boundary information instead of truncating the designated distribution and uses the boundary information to determine the similarity between observations and missings.

This paper consists of five sections. Our new imputation method is described and its properties discussed in Section 2. By simulation studies in Section 3, we compare our method with T-NORM and MV, PMM, and LRD with additional truncation procedure to meet boundary constraints and PRD-series methods to examine the effect of the double hot-deck step of our method, DBM-PRD. In Section 4, we apply DBM-PRD and the existing imputation methods to the 2013 health screening data of Korea for missing values of smoking periods. Finally, a brief conclusion is found in Section 5.

2 Double hot-deck boundary information matching proportioned residual draw

Suppose data are composed of a fully observed explanatory variable vector $\mathbf{X}_i$ and response variable Y_i for $i = 1, \ldots, n,$ for which some of Y_i's are missing (i.e., item missingness). Regardless of the missingness, let Y_i be individually bounded and values of the boundaries be given by

$$C_{i,L} \leq Y_i \leq C_{i,U} \tag{2.1}$$

where $C_{i,U}$ and $C_{i,L}$ are the upper and lower boundaries of $Y_i,$ $i = 1, \ldots, n_0, n_0 + 1, \ldots, n,$ and the first n_0 Y_i's are observed and remaining $(n - n_0)$ Y_i's are missing.

Following Rubin (1987), we generate the regression coefficients β^* and variance σ^{*2} from the posterior distributions given by

$$\sigma^{*2} \sim \hat{\sigma}^2_{\text{OLS}} \left(n_{\text{obs}} - q\right) / \chi^2_{n_{\text{obs}}-1}, \;\; \beta^* \sim N\left(\hat{\beta}_{\text{OLS}}, \sigma^{*2} \left(X^T X\right)^{-1}\right) \tag{2.2}$$

where X is the fully observed q covariates, and $\hat{\beta}_{\text{OLS}}$ and $\hat{\sigma}^2_{\text{OLS}}$ are the OLS estimates of regression coefficients and variance, respectively, from the regression model fitted to the observations. We then obtain the predictive means denoted by $\hat{Y}_i^{\text{obs}}$ for observed Y_i and $\hat{Y}_j^{\text{miss}}$ for missing Y_j. Then each $\hat{Y}_i^{\text{obs}}$ or $\hat{Y}_j^{\text{miss}}$ is located in one of following three intervals S^k where $k = -, 0, +$:

$$S^- = \left(-\infty, C_{iL}\right), \quad S^0 = \left(C_{iL}, C_{iU}\right), \quad S^+ = \left(C_{iU}, \infty\right).$$

For observed Y_i (i.e., $i = 1, \ldots, n_0$), we define the upper and lower proportioned residuals $\tilde{r}_{i,U}$ and $\tilde{r}_{i,L}$:

$$\tilde{r}_{i,U} = \frac{Y_i - \hat{Y}_i^{\text{obs}}}{C_{iU} - \hat{Y}_i^{\text{obs}}} \text{ and } \tilde{r}_{i,L} = \frac{Y_i - \hat{Y}_i^{\text{obs}}}{C_{iL} - \hat{Y}_i^{\text{obs}}} \tag{2.3}$$

where we assume that there is no $\hat{Y}$ which is exactly equal to its own upper or lower boundary.

These $\tilde{r}_{i,U}$ and $\tilde{r}_{i,L}$ in equation (2.3) are then both divided into three sets based on the S^k to which $\hat{Y}^{\text{obs}}$ belongs. For $k = -, 0, +,$

$$R_U^k = \left\{\tilde{r}_{i,U};\ \hat{Y}_i^{\text{obs}} \in S^k\right\} \text{ and } R_L^k = \left\{\tilde{r}_{i,L};\ \hat{Y}_i^{\text{obs}} \in S^k\right\}.$$

Finally, we impute the missing Y_j for $j = n_0 + 1, \ldots, n$ with

$$Y_{j,U}^* = \hat{Y}_j^{\text{miss}} + \tilde{r}_{j,U}^*\left(C_{j,U} - \hat{Y}_j^{\text{miss}}\right) \text{ or } Y_{j,L}^* = \hat{Y}_j^{\text{miss}} + \tilde{r}_{j,L}^*\left(C_{j,L} - \hat{Y}_j^{\text{miss}}\right). \tag{2.4}$$

In order to select $\tilde{r}_{j,U}^*$ or $\tilde{r}_{j,L}^*$ in (2.4), we now employ a hot-deck method that considers similar values as their candidates (i.e., possible donors). Hot-deck is a method for handling missing data in which each missing value is replaced with an observed response randomly selected from a donor containing similar units (Andridge and Little, 2010). We employ the following double hot-deck scheme.

1. [**The first hot-deck**] If $\hat{Y}_j^{\text{miss}} \in S^k$ for $k = -, 0, +$ and $\hat{Y}_j^{\text{miss}}$ is closer to C_{jU} than C_{jL}, then we select the corresponding upper proportioned residual set R_U^k as the set of possible donors for sampling $\tilde{r}_{jU}^*$. Likewise, if $\hat{Y}_j^{\text{miss}}$ is closer to C_{jL}, we select R_L^k for sampling $\tilde{r}_{jL}^*$.
2. [**The second hot-deck**] We construct possible donors from the R_U^k or R_L^k selected in the first hot-deck. The possible donors for sampling $\tilde{r}_{jU}^*$ consist of $\tilde{r}_{i,U}$'s whose $C_{iU} - \hat{Y}_i^{\text{obs}}$ is close to $C_{jU} - \hat{Y}_j^{\text{miss}}$ for R_U^k. Similarly, possible donors of $\tilde{r}_{i,L}$'s whose $C_{iL} - \hat{Y}_i^{\text{obs}}$ is close to $C_{jL} - \hat{Y}_j^{\text{miss}}$ for R_L^k.
3. [**Imputing**] Then $\tilde{r}_{i,U}$ or $\tilde{r}_{i,L}$ is randomly sampled from the corresponding possible donors to impute missing Y_j with $Y_{j,U}^*$ or $Y_{j,L}^*$, respectively and the selected $\tilde{r}_{i,U}$ and $\tilde{r}_{i,L}$ for $i = 1, \ldots, n_0$ are denoted by $\tilde{r}_{j,U}^*$ or $\tilde{r}_{j,L}^*$ for $j = n_0 + 1, \ldots, n$. Here the cases with $Y_{j,U}^* \leq C_{j,L}$ and/or $Y_{j,L}^* \geq C_{j,U}$ are excluded from the possible donor set. This is rare, although it does occur.

Theorem 1 *The values* $Y_{j,U}^*$ *and* $Y_{j,L}^*$ *always satisfy their boundary conditions.*

$$C_{j,L} \leq Y_{j,U}^* \leq C_{j,U} \text{ and } C_{j,L} \leq Y_{j,L}^* \leq C_{j,U}.$$

The proof is given in the Appendix.

Theorem 1 states that the boundary conditions of Y_j for $j = n_0 + 1, \ldots, n$ are always satisfied as DBM-PRD imputes missing Y_j with $Y_{j,U}^*$ or $Y_{j,L}^*$. We may assume that there exists only an upper boundary value such that $C_{i,L} = -\infty$ for $i = 1, \ldots, n$. Then the first hot-deck is not needed because R_U^k is automatically selected and $Y_{j,U}^* \geq C_{j,L} = -\infty$.

Corollary 1.1 *The DBM-PRD method is reduced to the boundary information matching method in Kwon and Park (2015), when there is only an upper or lower bound.*

To examine the double hot-deck procedures used in DBM-PRD, we consider three variations of DBM-PRD. The first variation is a *proportioned residual draw method* (PRD) which removes the two hot-decksteps from DBM-PRD and the second variation removes the first hot-deck procedure denoted by SPRD. Thus in PRD we randomly sample from all elements in R_U^k and R_L^k. In SPRD, the possible donor set is based on the minimum distance from either boundary to the predictive mean. Among possible donors for $\tilde{r}_{j,U}^*$ and $\tilde{r}_{j,L}^*$, we select and construct final donors based solely on the distance order, without distinction between upper and lower bounds. The third variation, denoted by SPRD_2, also removes the first hot-deck step as in SPRD and additionally changes the matching method in the second hot-deck step. The possible donor in SPRD_2 consists of $\tilde{r}_{i,U}$ and $\tilde{r}_{i,L}$ whose predicted mean $\hat{Y}_i^{\text{obs}}$ is close to $\hat{Y}_j^{\text{miss}}$.

3 Simulation

We use the following abbreviations for imputation methods discussed in Section 1 and 2; OBS (available cases), T-NORM (truncated normal imputation in Rubin (1978); Raghunathan et al. (2001)), MV (method adjusted for uncertainty of the mean and variance in Rubin and Schenker (1986)), PMM (predictive mean matching in Little (1988)), LRD (local residual draw method in Schenker et al. (2006)) and three variations of DBM-PRD denoted by PRD, SPRD and SPRD_2. We compare these eight imputation methods with our DBM-PRD where a truncation procedure is added in MV, PMM, and LRD to accommodate the boundary constraints, denoted by T-MV, T-PMM, and T-LRD, respectively.

We consider a sample size of 1,000 with a 20% or 50% missing rates from the following linear model:

$$Y_i = X_i + \varepsilon_i, \text{ where } C_{iL} \le Y_i \le C_{iU} \text{ for } i = 1, \ldots, n, \tag{3.1}$$

and X_i's are independently generated from $N(2, 2)$ and i.i.d. ε_i are simulated from $N(0, \sigma_Y)$ or the t-distribution with degree of freedom t_{df}. The boundary values $C_{i,U}$ and $C_{i,L}$ are generated with $Y_i + |Z_{i,U}|$ and $Y_i - |Z_{i,L}|$ where $Z_{i,U} \sim N(0, \sigma_U)$ and $Z_{i,L} \sim N(0, \sigma_L)$, respectively. We set $\text{Cor}(X, Y)$ to be 0.7 or 0.9 by adjusting σ_Y (or t_{df}), and $\text{Cor}(Y, C_U)$ and $\text{Cor}(Y, C_L)$ to be between 0 and 0.9 by adjusting σ_U and σ_L. The correlation $\text{Cor}(Y, C_U)$ $(\text{Cor}(Y, C_L))$ denoted by ρ_{y,c_u} (ρ_{y,c_l}) indicates that the upper bound C_U has stronger information for Y than the lower bound C_L when ρ_{y,c_u} is greater than ρ_{y,c_L} in absolute value.

Two types of missing mechanisms are considered. First, 20% of Y values are randomly chosen and treated as missing to reflect the "missing completely at random (MCAR)" missing mechanism. Second, we set 80% of Y_i's to missing when the corresponding X_i is greater than its mean and 20% of Y_i's to missing when the corresponding X_i is less than its mean. This results in approximately 50% of Y_i's with missing values overall and reflects "missing at random (MAR)". Note that no imputation is needed for missing values under MCAR, while imputation for missing values under MAR is required (Scheffer, 2002).

We repeat each simulation scenario 1,000 times with the number of imputations M equal to 5 and the number of possible donors in the selection pool for imputation m_d equal to 6. A possible donor size m_d is allowed to be smaller than 6 when there is not enough sample to compose a donor, but there is no such case when the sample size is 1,000. We choose the commonly used fixed numbers $M = 5$ and $m_d = 6$ (Geraci and McLain, 2018; Schafer, Ezzati-Rice, Johnson, Khare, Little and Rubin, 1996; Schenker and Taylor, 1996), because it is known that such a setup does not affect the performance of imputation methods significantly as shown in Schafer (1999) and Schenker and Taylor (1996).

The imputation methods are compared in terms of estimation accuracy and efficiency for population quantities: mean (μ) and the 5th, 25th, 50th, 75th, and 95th percentiles. Statistical inference after multiple imputation proceeds as in Rubin (2004) and Schafer et al. (1996). We use the mean absolute error (MAE), root mean squared error (RMSE), a coverage rate of 95% confidence interval (CR) and an average width of 95% confidence interval (AWCI) as evaluation criteria for measuring the estimation accuracy and efficiency (Yucel and Demirtas, 2010; Yucel, He and Zaslavsky, 2008; Gelman, Van Mechelen, Verbeke, Heitjan and Meulders, 2005).

3.1 Simulation results under MCAR

Figure 3.1 shows the distribution of $\hat{\mu} - \mu$ (bias) in 1,000 simulated data sets with 20% of MCAR missing values under $\left(\rho_{y,c_l}, \rho_{y,c_u}\right) = (0.8, 0)$ and $\rho_{x,y} = 0.7.$ Since no imputation is necessary for missing values under MCAR in the estimation of mean and variance of $Y,$ OBS is unbiased for the mean of $Y,$ as expected. However, Figure 3.1 shows that all imputation methods, except for DBM-PRD, reveal an over-estimation problem. Observe that the lower boundary has strong information for $Y\,(\rho_{y,c_l} = 0.8)$ but the upper boundary has no information $(\rho_{y,c_u} = 0).$ Except for OBS and DBM-PRD, this asymmetric boundary information pushes up imputed values in the other imputation methods. To see the effect of asymmetric boundary information on imputation accuracy, different values of $\left(\rho_{y,c_l}, \rho_{y,c_u}\right)$ are considered in Table 3.1.

When upper and lower boundaries provide boundary information for Y in a symmetric way (i.e., $\left(\rho_{y,c_l}, \rho_{y,c_u}\right) = (0.9, 0.9))$ all imputation methods are comparable and are competitive with OBS. However, in the presence of asymmetric boundary information $(\rho_{y,c_l}, \rho_{y,c_u}) = (0.8, 0)$ or $(\rho_{y,c_l}, \rho_{y,c_u}) =$ $(0.5, 0.8),$ the estimation accuracy of the existing T-NORM, T-MV, T-PMM, and T-LRD is much worse than OBS and DBM-PRD. In particular, the coverage rate of 95% CIs (CR) is dramatically decreased as the degree of asymmetry increases. On the other hand, those of the PRD series (i.e., PRD, SPRD, DBM-PRD) are resistant to such asymmetry, indicating that the proportioned residual draw is resistant to asymmetric boundary information. Among the PRD series, DBM-PRD outperforms PRD and SPRD and is even better than OBS in terms of MAE and RMSE.

Notice that, except OBS and DBM-PRD, the imputed values by all other imputation methods make the distribution of Y lean toward the boundary with weaker boundary information. More precisely, all the imputation methods except OBS and DBM-PRD tend to over-estimate the true mean of $Y\,(E\,(Y) = 2)$ for $\left(\rho_{y,c_l}, \rho_{y,c_u}\right) = (0.8, 0)$ because $\rho_{y,c_u} < \rho_{y,c_l},$ whereas they tend to under-estimate the true mean for $\left(\rho_{y,c_l}, \rho_{y,c_u}\right) = (0.5, 0.8)$ because $\rho_{y,c_u} > \rho_{y,c_l}.$ This dependency is also observed with the MAR missing mechanism as discussed in the following section.

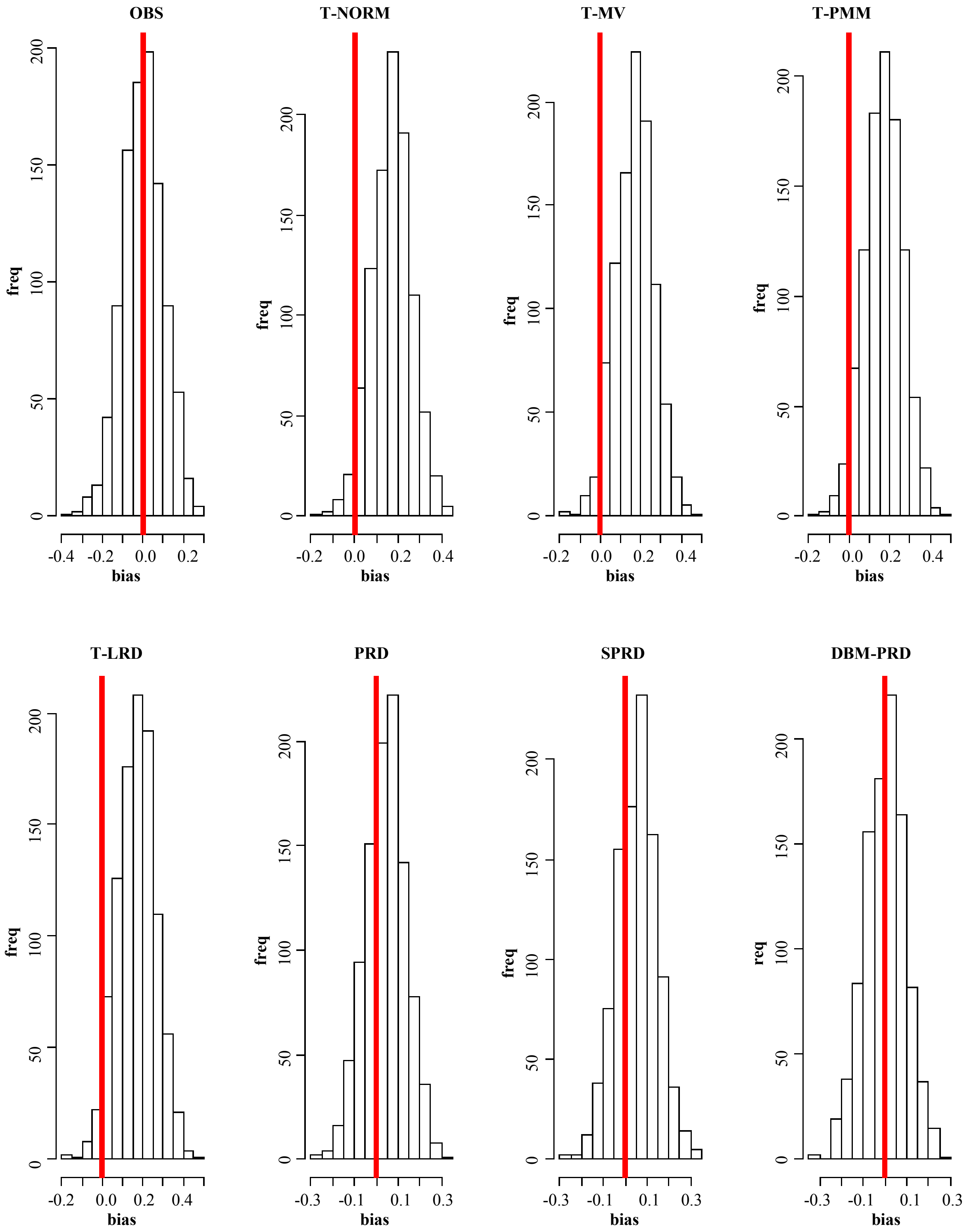

Figure 3.1 Distribution of bias, $\hat{\mu} - \mu$ in mean estimation with 20% MCAR missing values with normal error and $\left(\rho_{y,c_l}, \rho_{y,c_u}\right) = (0.8, 0)$ and $\rho_{x,y} = 0.7$.

Table 3.1
Simulation results of mean estimation $(\mu = 2)$ with 20% MCAR missing values with normal error

$\rho_{x,y}$	$(\rho_{y,c_l}, \rho_{y,c_u})$		OBS	T-NORM	T-MV	T-PMM	T-LRD	PRD	T-PRD	SPRD	DBM-PRD
0.9	(0.9, 0.9)	$\hat{\mu}$	2.003	2.003	2.003	2.003	2.003	2.003	2.003	2.003	2.003
		MAE	0.064	0.056	0.057	0.057	0.057	0.057	0.057	0.057	0.057
		RMSE	0.081	0.071	0.071	0.071	0.072	0.071	0.071	0.071	0.071
		CR (%)	94.9	95.8	95.3	95.5	94.8	95.6	95.2	95.5	95.2
		AWCI	0.310	0.280	0.280	0.279	0.279	0.283	0.279	0.279	0.278
0.7	(0.8, 0)	$\hat{\mu}$	2.000	2.171	2.171	2.171	2.170	2.043	2.044	2.055	2.000
		MAE	0.080	0.174	0.174	0.174	0.173	0.083	0.084	0.088	0.075
		RMSE	0.101	0.194	0.195	0.195	0.194	0.103	0.103	0.109	0.094
		CR (%)	94.9	54.1	54.3	52.1	53.7	92.3	91.4	89.8	94.2
		AWCI	0.393	0.367	0.366	0.362	0.363	0.362	0.358	0.358	0.359
0.7	(0.5, 0.8)	$\hat{\mu}$	2.000	1.906	1.906	1.906	1.907	1.927	1.973	1.979	1.983
		MAE	0.080	0.108	0.109	0.109	0.109	0.096	0.076	0.075	0.074
		RMSE	0.102	0.131	0.132	0.132	0.132	0.118	0.096	0.095	0.094
		CR (%)	94.2	83.0	83.7	83.0	82.6	88.7	93.3	93.7	94.0
		AWCI	0.393	0.362	0.361	0.359	0.359	0.379	0.358	0.356	0.355

3.2 Simulation results under MAR

Table 3.2 summarizes the results with 50% MAR missing values under normal and t $(t_{df} = 3)$ distribution errors. As expected, OBS which uses only observed values in estimation is much worse than any imputation method for estimating the true mean $\mu = 2$. The results related to asymmetric boundary information are along the same lines as the MCAR simulation results. The accuracy and efficiency of T-NORM, T-MV, T-PMM, and T-LRD are much worse than those of the PRD series when the boundary information is asymmetric. This shows the effect of the two hot-deck steps and the proportioned residual draw on the accuracy and efficiency of mean estimation. Except for the symmetric boundary information under normal and t distributions, the CRs of T-NORM, T-MV, T-PMM and T-LRD are less than 50%, much smaller than the target 95%. All imputation methods except DBM-PRD produce the empirical distribution of Y biased to the boundary with weaker boundary information under both normal and t error distributions. This implies that only DBM-PRD is resistant to asymmetric boundary information and error distributions regardless of the missing mechanism. As a result, DBM-PRD outperforms the other imputation methods in all simulation scenarios.

In order to see the effect of the double hot-deck procedures employed in DBM-PRD, we compared DBM-PRD with SPRD. The effect of the first hot-deck step can be examined by comparing DBM-PRD and SPRD as the first hot-deck step of DBM-PRD is removed in SPRD. DBM-PRD is consistently better than SPRD regardless of the evaluation measure as long as the boundary information is asymmetric. The CR of SPRD relative to that of DBM-PRD, becomes worse as the boundary information becomes more skewed to one side. Thus, the first hot-deck step has an important role in resisting asymmetry of the boundary information.

The role of the second hot-deck step can be checked by comparing SPRD and PRD where PRD does not adopt both hot-deck steps. SPRD is better than PRD when the boundary information is moderately asymmetric in normal error or t distributed error, implying that the second hot-deck step works for a heavier tail distribution than normal. This comparison is further discussed in the following evaluation for percentile estimation.

As we described before, the SPRD_2 is the same as SPRD except for the method of matching to construct possible donors. The possible donor in SPRD_2 consists of $\tilde{r}_{i,U}$ and $\tilde{r}_{i,L}$ whose predicted mean $\hat{Y}_i^{\text{obs}}$ is close to $\hat{Y}_j^{\text{miss}}$. Thus, by comparing SPRD and SPRD_2, we examine the effect of the boundary information matching used in DBM-PRD. Table 3.2 shows that SPRD outperforms SPRD_2 regardless of boundary information and error distributions, implying that the boundary information matching works better than the usual mean matching for imputation of bounded missing data.

Table 3.2
Simulation results of mean estimation $(\mu = 2)$ when 50% MAR

$(\rho_{x,y}, \rho_{y,c_l}, \rho_{y,c_u})$		OBS	T-NORM	T-MV	T-PMM	T-LRD	PRD	SPRD	SPRD_2	DBM-PRD
					normal distributed error					
(0.9, 0.9, 0.9)	$\hat{\mu}$	1.045	2.002	2.002	1.995	2.002	2.001	2.001	2.000	2.001
	MAE	0.955	0.057	0.059	0.058	0.059	0.059	0.059	0.06	0.06
	RMSE	0.958	0.072	0.075	0.074	0.074	0.074	0.075	0.076	0.075
	CR (%)	0.0	95.3	94.0	94.5	94.1	94.6	93.8	93.2	93.7
	AWCI	0.355	0.291	0.287	0.282	0.283	0.294	0.285	0.29	0.285
(0.7, 0.8, 0)	$\hat{\mu}$	1.043	2.426	2.425	2.417	2.424	2.097	2.103	2.098	1.993
	MAE	0.957	0.426	0.425	0.417	0.424	0.12	0.123	0.124	0.088
	RMSE	0.963	0.44	0.442	0.434	0.441	0.148	0.15	0.152	0.109
	CR (%)	0.0	4.0	5.2	3.6	3.6	80.5	77.7	78.5	92.0
	AWCI	0.467	0.454	0.428	0.393	0.395	0.398	0.379	0.382	0.38
(0.7, 0.5, 0.8)	$\hat{\mu}$	1.045	1.77	1.771	1.761	1.771	1.822	1.952	1.828	1.961
	MAE	0.955	0.231	0.23	0.239	0.229	0.182	0.091	0.177	0.087
	RMSE	0.961	0.249	0.251	0.259	0.249	0.205	0.114	0.203	0.11
	CR (%)	0.0	36.4	36.5	27.5	31.7	62.3	90.0	62.7	90.3
	AWCI	0.467	0.396	0.379	0.361	0.362	0.42	0.375	0.408	0.375
(0.55, 0.5, 0.5)	$\hat{\mu}$	1.041	1.992	1.990	1.983	1.991	1.989	1.991	1.990	1.991
	MAE	0.959	0.110	0.124	0.118	0.118	0.123	0.123	0.131	0.124
	RMSE	0.971	0.138	0.155	0.148	0.149	0.155	0.155	0.165	0.156
	CR (%)	0.0	95.6	89.9	88.7	89.0	93.1	89.4	89.3	89.4
	AWCI	0.611	0.557	0.520	0.486	0.488	0.584	0.508	0.555	0.513
(0.55, 0.5, 0.8)	$\hat{\mu}$	1.042	1.613	1.613	1.604	1.613	1.754	1.902	1.761	1.906
	MAE	0.958	0.387	0.387	0.396	0.387	0.249	0.128	0.243	0.126
	RMSE	0.969	0.404	0.406	0.414	0.406	0.276	0.158	0.271	0.156
	CR (%)	0.0	11.3	11.2	8.1	9.4	56.1	85.8	53.4	86.5
	AWCI	0.610	0.495	0.480	0.460	0.461	0.514	0.467	0.504	0.465
					t(3) distributed error					
(0.77, 0.9, 0.9)	$\hat{\mu}$	1.049	2.005	2.005	2.000	2.005	2.004	2.005	2.004	2.005
	MAE	0.951	0.067	0.069	0.067	0.067	0.069	0.069	0.069	0.07
	RMSE	0.956	0.084	0.087	0.084	0.084	0.087	0.087	0.086	0.087
	CR (%)	0.0	96.2	95.2	95.5	96.0	95.5	95.1	95.5	95.4
	AWCI	0.428	0.341	0.335	0.328	0.329	0.341	0.333	0.336	0.334
(0.77, 0.9, 0)	$\hat{\mu}$	1.042	2.401	2.357	2.354	2.356	2.128	2.091	2.12	2.005
	MAE	0.958	0.401	0.357	0.354	0.356	0.138	0.108	0.131	0.076
	RMSE	0.964	0.421	0.375	0.374	0.375	0.165	0.133	0.156	0.097
	CR (%)	0.0	2.7	6.5	4.9	5.2	71.3	80.1	72.3	93.2
	AWCI	0.429	0.407	0.391	0.37	0.37	0.36	0.338	0.348	0.342
(0.77, 0.5, 0.9)	$\hat{\mu}$	1.045	1.776	1.818	1.809	1.817	1.866	1.955	1.872	1.963
	MAE	0.955	0.224	0.185	0.192	0.185	0.142	0.086	0.137	0.083
	RMSE	0.961	0.244	0.207	0.212	0.205	0.165	0.107	0.161	0.104
	CR (%)	0.0	31.7	44.0	37.1	42.1	67.7	87.6	67.3	88.7
	AWCI	0.431	0.355	0.342	0.326	0.329	0.359	0.338	0.351	0.338

We also investigate the percentile estimation by evaluating how well each imputation method estimates the probability that Y is greater than 5%, 25%, 50%, 75%, 95% quantiles. Denote the p^{th} quantile by $y^{-1}(p)$ satisfying $P(Y > y^{-1}(p)) = 1 - p$. The five percentiles are chosen to test how different the true distribution of Y and the estimated distribution of Y are for different imputation methods. Table 3.3 shows the results of percentile estimation when 50% missing values under MAR with normal error when $(\rho_{x,y}, \rho_{y,c_l}, \rho_{y,c_u}) = (0.7, 0.5, 0.8)$ and $(0.7, 0.8, 0)$. From the first line of each table, the existing methods clearly produce distributions skewed to the right when $(\rho_{x,y}, \rho_{y,c_l}, \rho_{y,c_u}) = (0.7, 0.5, 0.8)$ because $\rho_{y,c_u} > \rho_{y,c_l}$, while they produce distributions skewed to the left when $(\rho_{x,y}, \rho_{y,c_l}, \rho_{y,c_u}) = (0.7, 0.8, 0)$ because $\rho_{y,c_u} < \rho_{y,c_l}$.

Table 3.3
Simulation results of percentile (P_k, where $P_k = P(Y \ge y^{-1}(k))$ and $y^{-1}(p)$ satisfying $P(Y > y^{-1}(p)) = 1 - p$) estimation when 50% MAR missing with normal error

Criterion	Parameter	OBS	T-NORM	T-MV	T-PMM	T-LRD	PRD	SPRD	SPRD$_2$	DBM-PRD
					$(\rho_{x,y}, \rho_{y,c_l}, \rho_{y,c_u}) = (0.7, 0.5, 0.8)$					
mean	$P_{0.05}$	0.922	0.947	0.947	0.947	0.947	0.935	0.950	0.936	0.951
	$P_{0.25}$	0.637	0.732	0.732	0.731	0.732	0.733	0.750	0.734	0.752
	$P_{0.50}$	0.350	0.467	0.467	0.465	0.466	0.494	0.497	0.495	0.500
	$P_{0.75}$	0.139	0.216	0.216	0.216	0.217	0.232	0.240	0.233	0.241
	$P_{0.95}$	0.022	0.037	0.037	0.036	0.037	0.045	0.043	0.044	0.043
MAE	$P_{0.05}$	0.028	0.006	0.006	0.007	0.007	0.015	0.005	0.015	0.005
	$P_{0.25}$	0.113	0.019	0.019	0.021	0.021	0.018	0.011	0.018	0.011
	$P_{0.50}$	0.150	0.034	0.033	0.036	0.035	0.014	0.013	0.015	0.013
	$P_{0.75}$	0.111	0.034	0.034	0.035	0.034	0.020	0.015	0.021	0.014
	$P_{0.95}$	0.028	0.013	0.013	0.015	0.014	0.007	0.008	0.008	0.009
CR (%)	$P_{0.05}$	21.6	95.1	95.0	90.4	92.3	57.6	96.8	56.3	96.9
	$P_{0.25}$	0.0	81.9	80.3	72.9	73.7	82.5	96.3	82.4	96.6
	$P_{0.50}$	0.0	60.7	59.7	50.5	52.3	95.5	95.6	92.5	95.5
	$P_{0.75}$	0.0	51.9	47.1	43.4	45.4	83.3	93.0	77.8	92.2
	$P_{0.95}$	7.5	77.4	77.7	58.3	63.0	97.6	93.9	92.9	92.3
					$(\rho_{x,y}, \rho_{y,c_l}, \rho_{y,c_u}) = (0.7, 0.8, 0)$					
mean	$P_{0.05}$	0.921	0.956	0.956	0.956	0.956	0.952	0.953	0.952	0.950
	$P_{0.25}$	0.637	0.780	0.781	0.781	0.781	0.761	0.764	0.762	0.750
	$P_{0.50}$	0.350	0.558	0.558	0.559	0.559	0.519	0.519	0.519	0.500
	$P_{0.75}$	0.137	0.314	0.314	0.313	0.313	0.257	0.260	0.257	0.249
	$P_{0.95}$	0.021	0.076	0.076	0.074	0.075	0.055	0.051	0.055	0.049
MAE	$P_{0.05}$	0.029	0.007	0.007	0.007	0.007	0.005	0.006	0.006	0.006
	$P_{0.25}$	0.113	0.031	0.031	0.031	0.031	0.015	0.017	0.015	0.012
	$P_{0.50}$	0.150	0.058	0.058	0.059	0.059	0.022	0.021	0.023	0.014
	$P_{0.75}$	0.113	0.064	0.064	0.063	0.063	0.015	0.016	0.018	0.012
	$P_{0.95}$	0.029	0.026	0.026	0.025	0.026	0.007	0.006	0.009	0.006
CR (%)	$P_{0.05}$	21.8	90.5	90.0	87.8	87.6	97.9	95.2	96.5	96.4
	$P_{0.25}$	0.0	48.1	45.9	43.8	43.7	89.3	84.0	86.6	95.9
	$P_{0.50}$	0.0	8.4	9.6	11.4	11.1	82.9	80.6	75.8	95.8
	$P_{0.75}$	0.0	7.0	6.6	11.0	9.9	93.6	88.2	83.8	96.7
	$P_{0.95}$	6.3	33.8	31.9	35.8	31.9	93.4	96.0	86.8	97.4

The degree of skewness is considerably weakened in the PRD series where DBM-PRD shows the best performance in percentile estimation. When the boundary information is moderately asymmetric (i.e., $\left(\rho_{x,y}, \rho_{y,c_l}, \rho_{y,c_u}\right) = (0.7, 0.5, 0.8)$), the second hot-deck step is important whereas the first hot-deck step is less important because SPRD is better than PRD but is comparable to DBM-PRD. On the other hand, when $\left(\rho_{x,y}, \rho_{y,c_l}, \rho_{y,c_u}\right) = (0.7, 0.8, 0)$ the first hot-deck step is more important to choose a correct boundary due to the extremely asymmetric boundary information, and DBM-PRD is best among PRD series because only it contains the first hot-deck step. In addition, SPRD is better than SPRD_2 in percentile estimation. In summary, double hot-deck procedures including boundary information matching and proportioned residual draw are essential not only for boundary restrictions but also for asymmetric boundary information and even for symmetric boundary information and heavy tail distributions.

4 Empirical analysis

4.1 Data

Health insurance services in Korea are national and compulsory by law, and the data related to medical information for the entire Korean population are recorded in a national health information database. A sample cohort database is constructed by stratified random sampling from this national health information for research purposes. It maintains a cohort structure continued from a sample in 2002 (Lee, Lee, Park, Shin and Kim, 2016). Based on this recently published medical big data, several medical studies have been conducted (Kwon, Lim and Park, 2017; Kim, Kwon, Yu, Kim, Choi, Baik, Park and Kim, 2017; Kim, Lee, Kim, Kim, Choi, Baik, Choi, Pop-Busui, Park and Kim, 2015; Ko, Yoon, Kim, Kim, Kim and Seo, 2016; Ko, Jo, Park, Kim, Kim and Park, 2016; Rim, Kim, Han and Chung, 2015).

We apply imputation methods to this sample cohort data, in particular, to missing values of self-reported variables. In health screening records, there are variables measured by health-care professionals, such as height, weight, blood pressure, and blood sugar. They are reliable and completely observed. On the other hand, some health screening variables such as smoking period, exercise frequency, and drinking habits are self-reported. They are likely to be incomplete and inaccurate as discussed by Crossley and Kennedy (2002), Cambois, Robine and Mormiche (2007), and Kwon and Park (2016).

Using health screening data of the sample cohort data for 2011 and 2013, we impute the missing values of smoking periods (in years) of current smokers in 2013 whose ages were between 20 and 84 and who had health screening records both in 2011 and 2013. Let $Y_{k,i}^{\text{self}}$ be the self-reported smoking period and $\text{AGE}_{k,i}$ be the minimum value of age categorized in 5-year classes for person i in year k, respectively.

There is no missing data in $Y_{2013,i}^{\text{self}}$ as we limited our analysis to those who answered that they were current smokers in 2013. Unreasonable values of smoking periods in $Y_{2013,i}^{\text{self}}$ are treated as missing, by comparing smoking periods and ages in 2011 and 2013; i.e. if $Y_{2013,i}^{\text{self}}$ meets any of the conditions $Y_{2013,i}^{\text{self}} > \text{AGE}_{2013,i} - a$, $Y_{2013,i}^{\text{self}} < Y_{2011,i}^{\text{self}} + 2 - b_1$, or $Y_{2013,i}^{\text{self}} > Y_{2011,i}^{\text{self}} + 2 + b_2$ where a is a minimum age started smoking, and b_1 and b_2 are tolerances according to human memory. We denote by $Y_{2013,i}$ the new smoking period in 2013 to distinguish it from $Y_{2013,i}^{\text{self}}$ with no missing values.

As Raghunathan et al. (2001) did, we set the upper bound, C_{Ui}, to be $\text{AGE}_{2013,i}$ −minimum age started smoking. Although Raghunathan et al. (2001) defined the minimum smoking age to be 18 years old, we take the minimum age to be 10 which is the lowest smoking age observed in the sample cohort data. We set $Y_{2011,i}^{\text{self}}$ for a lower boundary C_{Li}. When $Y_{2011,i}^{\text{self}}$ is missing which is 0.09% of data, we set C_{Li} to 0. If $Y_{2011,i}^{\text{self}}$ is the same as $Y_{2013,i}$ then we set the C_{Li} to $Y_{2011,i}^{\text{self}} - 0.001$ to ensure that no denominator of the proportioned residual given in equation (2.3) is zero.

By adjusting b_1 and b_2 which show the extent to which we allow the error due to human memory, the missing rate ranges from 44.0% to 73.5%. By taking $b_1 = b_2 = 2$ when we may allow up to two years of human memory error, the rate of missing data is 44.0%. When we do not allow any error at all, that is, letting $b_1 = 1, b_2 = 1$ for those who smoked in 2013 and did not smoke in 2011 and letting $b_1 = 2, b_2 = 1$ for those who smoked both in 2013 and 2011, the rate of missing data is 73.5%.

In order to fit the regression model for smoking periods, we use sex, age and income level as siginificant predictors, Table 4.1 shows the summary of data we use in this paper. Individual income information used to estimate health insurance premiums was observed in the form of a categorized ordered variable with 11 levels. We re-categorized the income groups into 3 groups: high (top 30%), low (bottom 30%), and medium (others) as this improved the fit.

Table 4.1
Summary of data

		missing rate (%)	**n**	**mean age**	**male ratio (%)**	**income group ratio (%)**	
						top 30%	**bottom 30%**
with ± 2 years tolerance	obs. missing	44.0	19,601	42.6	95.8	47.6	10.4
			15,414	48.5	95.2	44.3	15.6
without tolerance	obs. missing	73.5	9,266	38.7	95.9	49.1	7.6
			25,749	47.6	95.4	45.1	14.5
	total		35,015	45.2	95.6	46.1	12.7
		$\mathbf{cor}\left(Y_{2013}, \text{AGE}_{2013}\right)$			$\mathbf{cor}\left(Y_{2013}^{\text{self}}, Y_{2011}^{\text{self}}\right)$		
with ± 2 years tolerance		0.79			0.99		
without tolerance		0.71			0.99		

The average age of people who did not respond to the smoking period question 2013 is about 6 to 9 years older than that of respondents. The distribution and average of the income indicate that the nonrespondent's income level is lower than the respondent's income level. Since age and income level are important predictors for smoking period, it is hard to assume that missing mechanism of the smoking period is MCAR.

As seen in Table 4.1, the correlation between $Y_{2013,i}$ and $Y_{2011,i}^{\text{self}}$ is as high as 0.99 because of treating Y_{2013}^{self} as a missing value when not satisfying the logical constraints under the assumption that Y_{2011}^{self} is correct. However, Y_{2011}^{self} of course has the same problem as Y_{2013}^{self} and it is not reliable either. Despite such a very high correlation, this is the reason for not including $Y_{2011,i}^{\text{self}}$ as a predictor. Erroneous boundary information only affects the individual imputed values, but erroneous boundary information as a predictor affects the overall regression model estimates, which greatly affects the overall reliability of imputation. Hence, we only use the measured variables such as sex, age, and income which are collected by the government as a basis for the collection of national medical insurance premiums. Note that the age variable is also used as upper boundary information.

Since it is highly possible that the self-reported smoking periods in 2011 are not correct, there should be a criticism for setting Y_{2011}^{self} to be a lower bound. Thus, we consider another lower bound with $C_{Li} = 1$ which is the observed smallest smoking period of current smokers.

4.2 Results

The regression model for the smoking period in 2013, $Y_{2013,i}$ is fitted with fully observed cases as given by Table 4.2.

In Table 4.2, sex(female) is a dummy variable with value 1 indicating female, age is the central value of age categorized in 5-year classes, and income(low) and income(mid) are both dummy variables with value 1 indicating membership to the particular income group.

Table 4.2
Regression model for the smoking period in 2013

	with ± 2 year tolerance		without tolerance	
	estimate	**t-value**	**estimate**	**t-value**
intercept	-9.42	-54.00	-6.70	-23.97
sex(female)	-8.38	-40.51	-8.58	-28.41
age	0.69	182.34	0.64	96.56
income(low)	-0.32	-2.25	-0.76	-3.21
income(mid)	-0.50	-5.77	-0.99	-7.81
R square	0.66		0.55	

Table 4.3 shows the mean of $Y_{2013,i}$ estimated by each of five imputation methods with $M = 5$ and $m_d = 6$. A possible donor size m_d is allowed to be smaller than 6 when there is not enough sample to compose a donor, but there is no such case in our data. We consider four different scenarios made up of two settings of lower boundaries and two tolerances of human memory errors.

Different from the simulation results, T-NORM under-estimates more seriously than OBS when $C_{Li} = 1$. The distribution of observed smoking periods is slightly skewed to the right as the distance between Q50 and Q95 is farther than that between Q50 and Q5. However, T-NORM imputes a predictive mean plus a random residual generated from a truncated normal distribution not from the empirical distribution of residuals which is right-skewed. Since the lower bound, $C_{Li} = 1,$ is far from the mean, the possibility of selecting a negative error is higher from the truncated normal than from the right skewed empirical distribution of residuals. This leads to the underestimation of T-NORM. All other imputation methods estimate the mean smoking period higher than OBS as they use empirical residuals.

OBS produces higher mean of smoking periods with tolerance than without tolerance as the regression coefficient of age is higher with tolerance than without tolerance as shown in Table 4.2. Except with T-NORM, the estimated smoking periods are longer without tolerance than with tolerance, and the gap is smaller when $C_{Li} = 1$ than $C_{Li} = Y_{2011,i}$.

The DBM-PRD method is the most robust regardless of how we define missing values and the lower boundary. This is a desirable property of imputation when the boundary information is unreliable. On the other hands, the estimation results of the existing imputation methods (i.e., T-Norm, T-MV, T-LRD, T-PMM) clearly depend on the choice of boundary and human memory tolerance. The estimated distributions of smoking period by the existing methods move substantially to the right when $C_{Li} = Y_{2011,i}$ relative to when $C_{Li} = 1$ since $Y_{2011,i}$ is a more informative boundary than the constant boundary. However, the distribution with DBM-PRD is only marginally changed for different boundaries.

Table 4.3
Estimated mean, 5, 25, 50, 75 and 95% quantiles of smoking years of Korean current smokers in 2013

	with ± 2 years tolerance due to human memory (missing rate = 44.0%)											
	$C_{Li} = Y_{2011,i}$						$C_{Li} = 1$					
	mean	**Q5**	**Q25**	**Q50**	**Q75**	**Q95**	**mean**	**Q5**	**Q25**	**Q50**	**Q75**	**Q95**
OBS	19.55	7.00	12.00	20.00	25.00	40.00	19.55	7.00	12.00	20.00	25.00	40.00
T-NORM	21.46	8.00	14.00	20.00	27.98	40.22	17.66	3.92	10.00	16.99	22.91	35.00
T-MV	22.79	9.00	15.00	20.05	30.00	41.80	21.45	7.01	14.45	20.00	28.17	40.00
T-LRD	22.63	9.00	15.00	20.00	30.00	40.20	21.32	7.00	14.00	20.00	30.00	40.00
T-PMM	22.64	9.00	15.00	20.00	30.00	40.53	21.31	7.00	14.00	20.00	30.00	40.00
DBM-PRD	20.61	6.31	13.00	20.00	26.90	40.00	21.33	7.00	14.00	20.00	30.00	40.00
	without any tolerance due to human memory (missing rate = 73.5%)											
	$C_{Li} = Y_{2011,i}$						$C_{Li} = 1$					
	mean	**Q5**	**Q25**	**Q50**	**Q75**	**Q95**	**mean**	**Q5**	**Q25**	**Q50**	**Q75**	**Q95**
OBS	17.25	5.00	11.00	16.00	22.00	32.00	17.25	5.00	11.00	16.00	22.00	32.00
T-NORM	21.92	8.00	14.77	20.66	28.00	40.75	14.61	2.70	8.13	13.64	20.00	30.00
T-MV	24.01	10.00	16.75	22.85	30.34	42.67	21.63	7.19	15.00	20.95	27.68	37.87
T-LRD	24.40	10.00	16.00	23.00	30.00	47.40	21.44	5.00	13.00	20.00	28.80	42.00
T-PMM	24.41	10.00	16.00	23.00	30.00	47.45	21.47	5.00	13.00	20.00	28.80	42.00
DBM-PRD	21.11	7.00	13.00	20.00	27.00	41.80	21.42	5.00	13.00	20.00	28.80	42.00

Figure 4.1 presents the kernel density estimates of smoking periods using OBS and DBM-PRD under two lower boundary settings and two tolerances of human memory errors. Imputation by DBM-PRD moves the distribution of smoking period to the right and spreads it widely, compared to the distribution constructed only by observations (OBS).

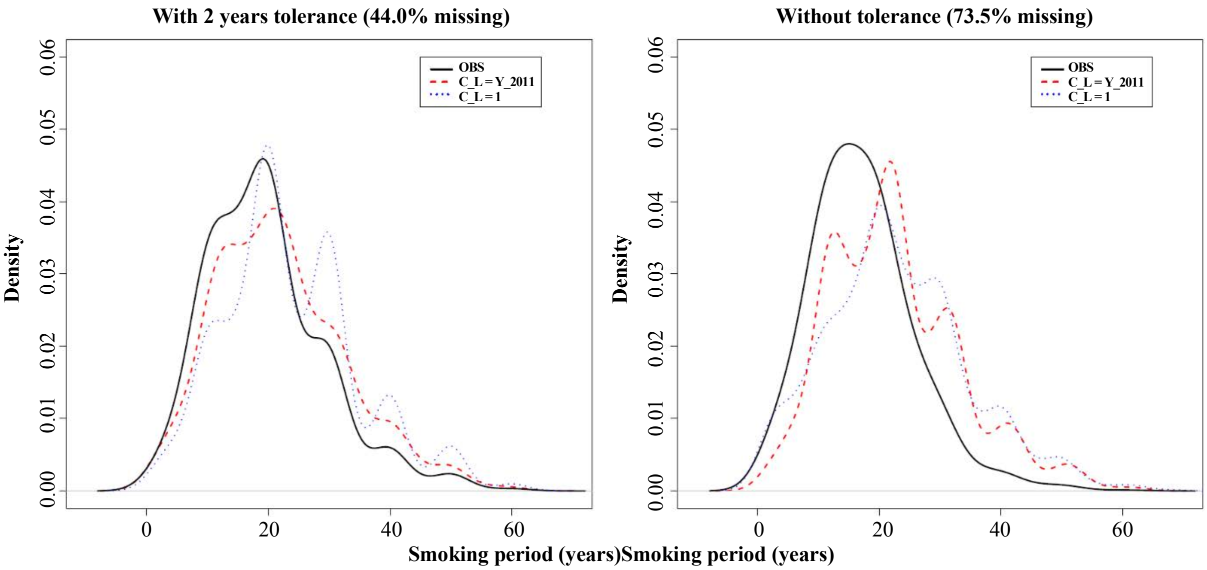

Figure 4.1 Kernel density estimates of smoking periods using OBS and DBM-PRD under two lower boundary settings and two tolerances of human memory errors.

5 Conclusion

We proposed a method for multiple imputation of missing variables when the missing values are logically bounded, which is often encountered in censuses or sample surveys. The existing imputation methods with an additional truncation or acceptance/rejection step produced biased estimates, depending on the extent of asymmetry of the boundary information. Their imputation values shrank toward the boundary with lower correlation with the missing variable. However, by employing a proportioned residual draw, boundary information matching, and a double hot-deck procedure, our DBM-PRD method produced more accurate and efficient estimates for the mean and percentiles, regardless of missingness rates, missing data mechanism, and distributions of the missing variable.

Moreover, our DBM-PRD imputation method is resistant to asymmetric boundary information in the sense that its imputed values do not depend on the extent of asymmetry of the boundary information. Especially, when there are two or more variables for the boundary information, or when reliability of the lower boundary information is suspected, DBM-PRD imputation is a powerful tool for estimating the parameters of interest accurately.

The DPM-PRD method also will work for a single imputation. There may be cases when (especially in official statistics) a single definitive output dataset is needed, and when users do not have the sophistication to deal with multiple imputation.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government Ministry of Science and ICT (MSIT) (NRF-2018R1C1B5043739). This work was supported by a Korea University Grant (K1910711) and Hankuk University of Foreign Studies Research Fund.

Appendix

Proof of Theorem 1

It suffices to show that $Y_{jU}^{*} \le C_{jU}$ and $Y_{jL}^{*} \ge C_{jL}$ because of the constraints in the imputation step.

1. If $\hat{Y}_j^{\text{miss}} \in S^0$, then $\tilde{r}_{jU}^{*}$ is sampled from R_U^0 whose element $\tilde{r}_{iU} \le 1$ for all i because $Y_i \le C_{iU}$ and $C_{iU} - \hat{Y}_i > 0$ for $\tilde{r}_{iU} \in R_U^0$. Since $\tilde{r}_{jU}^{*}$ is one of such $\tilde{r}_{iU}$'s, we have $\tilde{r}_{jU}^{*} \le 1$. Furthermore $C_{jU} - \hat{Y}_j^{\text{miss}} > 0$ gives

$$Y_{j,U}^{*} = \hat{Y}_j^{\text{miss}} + \tilde{r}_{j,U}^{*}\left(C_{j,U} - \hat{Y}_j^{\text{miss}}\right) \le \hat{Y}_j^{\text{miss}} + 1 \times \left(C_{j,U} - \hat{Y}_j^{\text{miss}}\right) = C_{jU}. \tag{A.1}$$

   Similarly, if $\hat{Y}_j^{\text{miss}} \in S^0$, then $\tilde{r}_{jL}^{*}$ is randomly selected from R_L^0 whose element $\tilde{r}_{iL} \le 1$ for any i because $Y_i \ge C_{iL}$ and $C_{iL} - \hat{Y}_i < 0$ for $i \in R_L^0$. Since $\tilde{r}_{jL}^{*}$ is one of such $\tilde{r}_{iL}$, $\tilde{r}_{jL}^{*} \le 1$. Using $C_{j,L} - \hat{Y}_j^{\text{miss}} < 0$ because $\hat{Y}_j^{\text{miss}} \in S^0$, we have

$$Y_{j,L}^{*} = \hat{Y}_j^{\text{miss}} + \tilde{r}_{jL}^{*}\left(C_{j,L} - \hat{Y}_j^{\text{miss}}\right) \ge \hat{Y}_j^{\text{miss}} + 1 \times \left(C_{j,L} - \hat{Y}_j^{\text{miss}}\right) = C_{jL}. \tag{A.2}$$

2. If $\hat{Y}_j^{\text{miss}} \in S^{+}$, then $\tilde{r}_{jU}^{*}$ is sampled from R_U^{+} whose element $\tilde{r}_{iU} \ge 1$ for all i because $Y_i \le C_{i,U}$ and $C_{iU} - \hat{Y}_i < 0$. Since $\tilde{r}_{jU}^{*}$ is one of such $\tilde{r}_{iU}$'s, we have $\tilde{r}_{jU}^{*} \ge 1$. Furthermore $C_{jU} - \hat{Y}_j^{\text{miss}} < 0$ gives

$$Y_{j,U}^{*} = \hat{Y}_j^{\text{miss}} + \tilde{r}_{j,U}^{*}\left(C_{j,U} - \hat{Y}_j^{\text{miss}}\right) \le \hat{Y}_j^{\text{miss}} + 1 \times \left(C_{j,U} - \hat{Y}_j^{\text{miss}}\right) = C_{jU}. \tag{A.3}$$

   Similarly, if $\hat{Y}_j^{\text{miss}} \in S^{+}$, then $\tilde{r}_{jL}^{*}$ is randomly selected from R_L^{+} whose element $\tilde{r}_{iL} \le 1$ for any i because $Y_i \ge C_{iL}$ and $C_{iL} - \hat{Y}_i < 0$ for $i \in R_L^{+}$. Since $\tilde{r}_{jL}^{*}$ is one of such $\tilde{r}_{iL}$, $\tilde{r}_{jL}^{*} \le 1$. Using $C_{j,L} - \hat{Y}_j^{\text{miss}} < 0$ because of $\hat{Y}_j^{\text{miss}} \in S^{+}$, we have

$$Y_{j,L}^{*} = \hat{Y}_j^{\text{miss}} + \tilde{r}_{jL}^{*}\left(C_{j,L} - \hat{Y}_j^{\text{miss}}\right) \ge \hat{Y}_j^{\text{miss}} + 1 \times \left(C_{j,L} - \hat{Y}_j^{\text{miss}}\right) = C_{jL}. \tag{A.4}$$

3. If $\hat{Y}_j^{\text{miss}} \in S^-$, then $\tilde{r}_{jU}^*$ is sampled from R_U^- whose element $\tilde{r}_{iU} \le 1$ for all i because $Y_i \le C_{iU}$ and $C_{iU} - \hat{Y}_i > 0$ for $\tilde{r}_{iU} \in R_U^-$. Since $\tilde{r}_{jU}^*$ is one of such $\tilde{r}_{iU}$'s, we have $\tilde{r}_{jU}^* \le 1$. Furthermore $C_{jU} - \hat{Y}_j^{\text{miss}} > 0$ gives

$$Y_{j,U}^* = \hat{Y}_j^{\text{miss}} + \tilde{r}_{j,U}^* \left(C_{j,U} - \hat{Y}_j^{\text{miss}} \right) \le \hat{Y}_j^{\text{miss}} + 1 \times \left(C_{j,U} - \hat{Y}_j^{\text{miss}} \right) = C_{jU}. \tag{A.5}$$

   Similarly, if $\hat{Y}_j^{\text{miss}} \in S^-$, then $\tilde{r}_{jL}^*$ is randomly selected from R_L^- whose element $\tilde{r}_{iL} \ge 1$ for any i because $Y_i \ge C_{iL}$ and $C_{iL} - \hat{Y}_i > 0$ for $i \in R_L^-$. Since $\tilde{r}_{jL}^*$ is one of such $\tilde{r}_{iL}$, $\tilde{r}_{jL}^* \ge 1$. Using $C_{j,L} - \hat{Y}_j^{\text{miss}} > 0$ because $\hat{Y}_j^{\text{miss}} \in S^-$, we have

$$Y_{j,L}^* = \hat{Y}_j^{\text{miss}} + \tilde{r}_{jL}^* \left(C_{j,L} - \hat{Y}_j^{\text{miss}} \right) \ge \hat{Y}_j^{\text{miss}} + 1 \times \left(C_{j,L} - \hat{Y}_j^{\text{miss}} \right) = C_{jL}. \tag{A.6}$$

References

Andridge, R.R., and Little, R.J.A. (2010). A review of hot-deck imputation for survey non-response. *International Statistical Review*, 78(1), 40-64.

Cambois, E., Robine, J.-M. and Mormiche, P. (2007). Did the prevalence of disability in France really fall in the 1990s? A discussion of questions asked in the French Health Survey. *Population-E*, 62(2), 313-337.

Crossley,T.F., and Kennedy, S. (2002). The reliability of self-assessed health status. *Journal of Health Economics*, 21, 4, 643-658.

Gelman, A., Van Mechelen, I., Verbeke, G., Heitjan, D.F. and Meulders, M. (2005). Multiple imputation for model checking: Completed-data plots with missing and latent data. *Biometrics*, 61(1), 74-85.

Geraci, M., and McLain, A. (2018). Multiple imputation for bounded variables. *Psychometrika*, 83(4), 919-940.

Honaker, J., King, G. and Blackwell, M. (2012). Amelia II: A program for missing data. R package, version 1.6.4.

Kim, N.H., Kwon, T.Y., Yu, S., Kim, N.H., Choi, K.M., Baik, S.H., Park, Y. and Kim, S.G. (2017). Increased vascular disease mortality risk in prediabetic Korean adults is mainly attributable to ischemic stroke. *Stroke*, 48, 4, 840-845.

Kim, N.H., Lee, J., Kim, T.J., Kim, N.H., Choi, K.M., Baik, S.H., Choi, D.S., Pop-Busui, R., Park, Y. and Kim, S.G. (2015). Body mass index and mortality in the general population and in subjects with chronic disease in Korea: A nationwide cohort study (2002-2010). *PLoS ONE*, 10(10).

Ko, M.J., Jo, A.J., Park, C.M., Kim, H.J., Kim, Y.J. and Park, D.W. (2016). Level of blood pressure control and cardiovascular events: SPRINT criteria versus the 2014 hypertension recommendations. *Journal of the American College of Cardiology*, 67(24), 2821-2831.

Ko, S., Yoon, S.J., Kim, D., Kim, A.R., Kim, E.J. and Seo, H.Y. (2016). Metabolic risk profile and cancer in Korean men and women. *Journal of Preventive Medicine and Public Health*, 49(3), 143-152.

Kwon, T.Y., and Park, Y. (2015). A new multiple imputation method for bounded missing values. *Statistics and Probability Letters*, 107, 204-209.

Kwon, T.Y., and Park, Y. (2016). Reliability of self-reported data for prevalence and health life expectancy studies: Comparison with sample cohort DB of National Health Insurance Services. *The Korean Journal of Applied Statistics*, 39(7), 1329-1346.

Kwon, T.Y., Lim, J. and Park, Y. (2017). Health life expectancy in Korea based on sample cohort database of National Health Insurance Services. *The Korean Journal of Applied Statistics*, 30(3), 475-486.

Lee, J., Lee, J.S., Park, S.H., Shin, S.A. and Kim, K. (2016). Cohort profile: The National Health Insurance Service-National Sample Cohort (NHIS-NSC), South Korea. *International Journal of Epidemiology*, doi: 10.1093/ije/dyv319.

Little, R. (1988). Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics*, 6, 287-296.

Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J. and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27, 1, 85-95. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2001001/article/5857-eng.pdf.

Raghunathan, T.E., Solenberger, P.W. and Van Hoewyk, J. (2002). Iveware: Imputation and variance estimation software. Ann Arbor, MI: Survey Methodology Program, Survey Research Center, Institute for Social Research, University of Michigan.

Rim, T.H., Kim, D.W., Han, J.S. and Chung, E.J. (2015). Retinal vein occlusion and the risk of stroke development: A 9-year nationwide population-based study. *Ophthalmology*, 122(6), 1187-1194.

Rubin, D.B. (1978). Multiple imputation in sample surveys - A phenomenological Bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section,* American Statistical Association, 20-34.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc.

Rubin, D.B. (2004). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc.

Rubin, D.B., and Schenker, N. (1986). Multiple imputations for interval estimation from simple random sampling with ignorable nonresponse. *Journal of American Statistical Association*, 81(394), 366-374.

Schafer, J.L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research*, 8(1), 3-15.

Schafer, J.L., Ezzati-Rice, T.M., Johnson, W., Khare, M., Little, R.J.A. and Rubin, D.B. (1996). The NHANES III multiple imputation project. *Race/Ethnicity*, 60(21.2).

Scheffer, J. (2002). Dealing with missing data. *Research Letters in the Information and Mathematical Sciences*, 3, 153-160.

Schenker, N., Raghunathan, T.E., Chiu, P.L., Makuc, D.M., Zhang, G. and Cohen, A.J. (2006). Multiple imputation of missing income data in the National Health Interview Survey. *Journal of the American Statistical Association*, 101(475), 924-933.

Schenker, N., and Taylor, J.M.G. (1996). Partially parametric techniques for multiple imputation. *Computational Statistics & Data Analysis*, 22(4), 425-446.

Su, Y.-S., Gelman, A., Hill, J. and Yajima, M.(2011). Multiple imputation with diagnostics (mi) in R: Opening windows into the black box. *Journal of Statistical Software*, 45(2), 1-31.

van Buuren, S., and Groothuis-Oudshoorn, K. (2010). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 1-67.

Yucel, R.M., and Demirtas, H. (2010). Impact of non-normal random effects on inference by multiple imputation: A simulation assessment. *Computational Statistics & Data Analysis*, 54(3), 790-801.

Yucel, R.M., He, Y. and Zaslavsky, A.M. (2008). Using calibration to improve rounding in imputation. *The American Statistician*, 62(2), 125-129.

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 35, No. 4, December 2019

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 36, No. 1, March 2020

All inquires about submissions and subscriptions should be directed to jos@scb.se

The Canadian Journal of Statistics **La revue canadienne de statistique**

CONTENTS TABLE DES MATIÈRES

Volume 47, No. 3, September/septembre 2019

The Canadian Journal of Statistics **La revue canadienne de statistique**

CONTENTS TABLE DES MATIÈRES

Volume 47, No. 4, December/décembre 2019

Original Articles

Review Article

Original Articles

GUIDELINES FOR MANUSCRIPTS

Authors are invited to submit their articles **through the *Survey Methodology* hub on the ScholarOne Manuscripts website** (https://mc04.manuscriptcentral.com/surveymeth). 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.