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

Survey Methodology

Survey Methodology
48-1

Release date: June 21, 2022

Statistics Canada Statistique Canada

Canada

How to obtain more information

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

You can also contact us by

Email at infostats@statcan.gc.ca

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

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 2022

•

Volume 48

•

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 usually publishes innovative theoretical or applied research papers, and sometimes review papers, that provide new insights on statistical methods relevant to National Statistical Offices and other statistical organizations. Topics of interest are provided on the journal web site (www.statcan.gc.ca/surveymethodology). Authors can submit papers either to the regular section of the Journal or to the short notes section for contributions under 3,000 words, including tables, figures and references. Although the review process may be streamlined for short notes, all papers are peer-reviewed. 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@statcan.gc.ca).

Survey Methodology

A Journal Published by Statistics Canada

Volume 48, Number 1, June 2022

Contents

Regular Papers

Survey Methodology, June 2022
Vol. 48, No. 1, pp. 1-23
Statistics Canada, Catalogue No. 12-001-X

Maximum entropy classification for record linkage

Danhyang Lee, Li-Chun Zhang and Jae Kwang Kim[1]

Abstract

By record linkage one joins records residing in separate files which are believed to be related to the same entity. In this paper we approach record linkage as a classification problem, and adapt the maximum entropy classification method in machine learning to record linkage, both in the supervised and unsupervised settings of machine learning. The set of links will be chosen according to the associated uncertainty. On the one hand, our framework overcomes some persistent theoretical flaws of the classical approach pioneered by Fellegi and Sunter (1969); on the other hand, the proposed algorithm is fully automatic, unlike the classical approach that generally requires clerical review to resolve the undecided cases.

Key Words: Probabilistic linkage; Density ratio; False link; Missing match; Survey sampling.

1. Introduction

Combining information from multiple sources of data is a frequently encountered problem in many disciplines. To combine information from different sources, one assumes that it is possible to identify the records associated with the same entity, which is not always the case in practice. The entity may be individual, company, crime, etc. If the data do not contain unique identification number, identifying records from the same entity becomes a challenging problem. *Record linkage* is the term describing the process of joining records that are believed to be related to the same entity. While record linkage may entail the linking of records within a single computer file to identify duplicate records, referred to as *deduplication*, we focus on linking of records across separate files.

Record linkage (RL) has been employed for several decades in survey sampling producing official statistics. In particular, linking administrative files with survey sample data can greatly improve the quality and resolution of the official statistics. As applications, Jaro (1989) and Winkler and Thibaudeau (1991) merged post-enumeration survey and census data for census coverage evaluation. Zhang and Campbell (2012) linked population census data files over time, and Owen, Jones and Ralphs (2015) linked administrative registers to create a single statistical population dataset. The classical approach pioneered by Fellegi and Sunter (1969), which is the most popular method of RL in practice, has been successfully employed for these applications.

The probabilistic decision rule of Fellegi and Sunter (1969) is based on the likelihood ratio test idea, by which we can determine how likely a particular record pair is a true match. In applying the likelihood ratio test idea, one needs to estimate the model parameters of the underlying model and determine the thresholds of the decision rule. Winkler (1988) and Jaro (1989) treat the matching status as an unobserved variable and propose an EM algorithm for parameter estimation, which we shall refer to as the

1. Danhyang Lee, Department of Information Systems, Statistics and Management Science, University of Alabama, Tuscaloosa, AL, U.S.A.; Li-Chun Zhang, Department of Social Statistics and Demography, University of Southampton, Southampton, U.K., Statistics Norway, Oslo, Norway and Department of Mathematics, University of Oslo, Oslo, Norway. E-mail: L.Zhang@soton.ac.uk; Jae Kwang Kim, Department of Statistics, Iowa State University, Ames, IA, U.S.A.

WJ-procedure. See Herzog, Scheuren and Winkler (2007), Christen (2012) and Binette and Steorts (2020) for overviews. However, as explained in Section 2, to motivate the WJ-procedure as an EM algorithm requires the crucial assumption that measures of agreement between the record pairs, called *comparison vectors*, are independent from one record pair to another, which is impossible to hold in reality. Newcombe, Kennedy, Axford and James (1959) address dependence between comparison vectors through data application. Also, see e.g. Tancredi and Liseo (2011), Sadinle (2017), and Binette and Steorts (2020) for discussions of this issue. Bayesian approaches to RL are also available in the literature (Steorts, 2015; Sadinle, 2017; Stringham, 2021). Bayesian approaches to RL problems allow us to quantify uncertainty on the matching decisions. However, the stochastic search using MCMC algorithm in the Bayesian approach involves extra computational burden.

To develop an alternative approach, we first note that the RL problem is essentially a classification problem, where each record pair is classified into either "match" or "non-match" class. Various classification techniques based on machine learning approaches have been employed for record linkage (Hand and Christen, 2018; Christen, 2012, 2008; Sarawagi and Bhamidipaty, 2002). In this paper, we adapt the maximum entropy method for classification to record linkage. Specifically, we can view the likelihood ratio of the method proposed by Fellegi and Sunter (1969) as a special case of the density ratio and apply the maximum entropy method for density ratio estimation. For example, Nigam, Lafferty and McCallum (1999) use the maximum entropy for text classification and Nguyen, Wainwright and Jordan (2010) develop a more unified theory of maximum entropy method for density ratio estimation. There is, however, a key difference of record linkage to the standard setting of classification problems, in that the different record pairs are not distinct 'units' because the same record is part of many record pairs.

We present our maximum entropy record linkage algorithm for both supervised and unsupervised settings, while our main contributions concern the unsupervised case. Supervised approaches need training data, i.e., record pairs with known true match and true non-match status. Such training data are often not available in real world situations, or have to be prepared manually, which is very expensive and time-consuming (Christen, 2007). Thus, the unsupervised case is by far the most common in practice. In the unsupervised case, however, one cannot estimate the density ratio directly based on the observed true matches and non-matches, and it is troublesome to jointly model for the unobserved match status and the observed comparison scores over all the record pairs. We develop a new iterative algorithm to jointly estimate the density ratio as well as the maximum entropy classification set in the unsupervised setting and prove its convergence. The associated measures of the linkage uncertainty are also developed.

Furthermore, we show that the WJ-procedure can be incorporated as a special case of our approach to estimation, but without the need of the independence assumption between the record pairs. This reveals that the WJ-procedure can be motivated without the independence assumption, and explains why it gives reasonable results in many situations. The choice of the set of links is guided by the uncertainty measures developed in this paper. This is an important practical improvement over the classical approach, which does not directly provide any uncertainty measure for the final set of links. Our procedure is fully

automatic, without the need for resource-demanding clerical review that is required under the classical approach.

The paper is organised as follows. In Section 2, the basic setup and the classical approach are introduced. In Section 3, the proposed method is developed under the setting of supervised record linkage. In Section 4, we extend the proposed method to the more challenging case of the unsupervised record linkage. Discussions of some related estimation approaches and technical details are presented in Section 5 and the supplementary material. Results from an extensive simulation study are presented in Section 6. Some concluding remarks and comments on further works are given in Section 7.

2. Problems with the classical approach

Suppose that we have two data files A and B that are believed to have many common entities but no duplicates within each file. Any record in A and another one in B may or may not refer to the same entity. Our goal is to find the true matches among all possible pairs of the two data files. Let the bipartite *comparison space* $\Omega = A \times B = M \cup U$ consist of *matches* M and *non-matches* U between the records in files A and B. For any pair of records $(a,b) \in \Omega$, let $\boldsymbol{\gamma}_{ab}$ be the *comparison vector* between a set of *key* variables associated with $a \in A$ and $b \in B$, respectively, such as name, sex, date of birth. The key variables and the comparison vector $\boldsymbol{\gamma}_{ab}$ are fully observed over Ω. In cases where the key variables may be affected by errors, a match (a,b) may not have complete agreement in terms of $\boldsymbol{\gamma}_{ab}$, and a non-match (a,b) can nevertheless agree on some (even all) of the key variables.

In the classical approach of Fellegi and Sunter (1969), one recognizes the probabilistic nature of $\boldsymbol{\gamma}_{ab}$ due to the perturbations that cause key-variable errors. The related methods are referred to as *probabilistic record linkage*. To explain the probabilistic record linkage method of Fellegi and Sunter (1969), let $m(\boldsymbol{\gamma}_{ab}) = f\left(\boldsymbol{\gamma}_{ab} \middle| (a,b) \in M\right)$ be the probability mass function of the discrete values $\boldsymbol{\gamma}_{ab}$ can take given $(a,b) \in M$. Similarly, we can define $u(\boldsymbol{\gamma}_{ab}) = f\left(\boldsymbol{\gamma}_{ab} \middle| (a,b) \in U\right)$. The ratio

$$r_{ab} = \frac{m(\boldsymbol{\gamma}_{ab})}{u(\boldsymbol{\gamma}_{ab})}$$

is then the basis of the likelihood ratio test (LRT) for $H_0\colon (a,b) \in M$ vs. $H_1\colon (a,b) \in U$. Let $M^* = \{(a,b)\colon r_{ab} > c_M\}$ be the pairs classified as matches and $U^* = \{(a,b)\colon r_{ab} < c_U\}$ the non-matches, the remaining pairs are classified by clerical review, where (c_M, c_U) are the thresholds related to the probabilities of false links (of pairs in U) and false non-links (of pairs in M), respectively, defined as

$$\mu = \sum_{\boldsymbol{\gamma}} u(\boldsymbol{\gamma})\, \delta(M^*;\, \boldsymbol{\gamma}) \quad \text{and} \quad \lambda = \sum_{\boldsymbol{\gamma}} m(\boldsymbol{\gamma})\, \delta(U^*;\, \boldsymbol{\gamma}), \tag{2.1}$$

where $\delta(M^*;\, \boldsymbol{\gamma}) = 1$ if $\boldsymbol{\gamma}_{ab} = \boldsymbol{\gamma}$ means $(a,b) \in M^*$ and 0 otherwise, similarly for $\delta(U^*;\, \boldsymbol{\gamma})$.

In practice the probabilities $m(\boldsymbol{\gamma})$ and $u(\boldsymbol{\gamma})$ are unknown. Neither is the *prevalence* of true matches, given by $\pi = |M|/|\Omega| := n_M/n$. Let $\boldsymbol{\eta}$ be the set containing π and the unknown parameters of $m(\boldsymbol{\gamma})$ and $u(\boldsymbol{\gamma})$. Let $g_{ab} = 1$ if $(a,b) \in M$ and 0 if $(a,b) \in U$. Given the complete data $\{(g_{ab}, \boldsymbol{\gamma}_{ab}) : (a,b) \in \Omega\}$, Winkler (1988) and Jaro (1989) assume the log-likelihood to be

$$h(\boldsymbol{\eta}) = \sum_{(a,b)\in\Omega} g_{ab} \log\left(\pi m(\boldsymbol{\gamma}_{ab})\right) + \sum_{(a,b)\in\Omega} (1 - g_{ab}) \log\left((1-\pi) u(\boldsymbol{\gamma}_{ab})\right). \tag{2.2}$$

An EM-algorithm follows by treating $g_\Omega = \{g_{ab} : (a,b) \in \Omega\}$ as the missing data.

There are two fundamental problems with this classical approach.

[Problem-I] Record linkage is not a direct application of the LRT, because one needs to evaluate *all* the pairs in Ω instead of any *given* pair. The classification of Ω into M^* and U^* is incoherent generally, since a given record can belong to multiple pairs in M^*. Post-classification deduplication of M^* would be necessary then, which is *not* part of the theoretical formulation above. In particular, there lacks an associated method for estimating the uncertainty surrounding the final linked set, such as the amount of false links in it or the remaining matches outside of it.

[Problem-II] In reality the comparison vectors of any two pairs are not independent, as long as they share a record. For example, given $(a,b) \in M$ and $\boldsymbol{\gamma}_{ab}$ not subjected to errors, then $g_{ab'}$ must be 0, for $b' \neq b$ and $b' \in B$, as long as there are no duplicated records in either A or B, and $\boldsymbol{\gamma}_{ab'}$ depends only on the key-variable errors of b'. Whereas, marginally, $g_{ab'} = 1$ with probability π and $\boldsymbol{\gamma}_{ab'}$ depends also on the key-variable errors of a. It follows that $h(\boldsymbol{\eta})$ in (2.2) does not correspond to the true joint-data distribution of $\boldsymbol{\gamma}_\Omega = \{\boldsymbol{\gamma}_{ab} : (a,b) \in \Omega\}$, even when the marginal m and u-probabilities are correctly specified. Similarly, although one may define *marginally* $\pi = \Pr\left[(a,b) \in M \,|\, (a,b) \in \Omega\right]$ for a *randomly* selected record pair from Ω, it does not follow that $\log f(g_\Omega) = n_M \log \pi + (n - n_M) \log(1-\pi)$ *jointly* as in (2.2). For both reasons, $h(\boldsymbol{\eta})$ given by (2.2) cannot be the complete-data log-likelihood.

In the next two sections, we develop maximum entropy classification to record linkage to avoid the problems above, after which more discussions of the classical approach will be given.

3. Maximum entropy classification: Supervised

As noted in Section 1, the record linkage problem is a classification problem. Maximum entropy classification has been used in image restoration or text analysis (Gull and Daniell, 1984; Berger, Della Pietra and Della Pietra, 1996). *Maximum entropy classification (MEC)* has been proposed for supervised learning (SL) to standard classification problems, where the units are known but the true

classes of the units are unknown apart from a sample of *labelled units*. Let $Y \in \{1, 0\}$ be the true class and $\mathbf{X}$ the random vector of features. Let the density ratio be

$$r(\mathbf{x}; \boldsymbol{\eta}) = \frac{f(\mathbf{x} \mid Y = 1; \boldsymbol{\eta})}{f(\mathbf{x} \mid Y = 0; \boldsymbol{\eta})} := \frac{f_1(\mathbf{x}; \boldsymbol{\eta})}{f_0(\mathbf{x}; \boldsymbol{\eta})},$$

where f_1 and f_0 are the conditional density functions given $Y = 1$ or 0, respectively, and $\boldsymbol{\eta}$ contains the unknown parameters. For MEC based on $r(\mathbf{x})$, one finds $\hat{\boldsymbol{\eta}}$ that maximises the Kullback-Leibler (KL) divergence from f_0 to f_1 subjected to a constraint, i.e.

$$D = \int_{S_1} f_1(\mathbf{x}; \boldsymbol{\eta}) \log r(\mathbf{x}; \boldsymbol{\eta})\, d\mathbf{x} \quad \text{subjected to} \quad \int_{S_1} f_0(\mathbf{x}; \hat{\boldsymbol{\eta}})\, r(\mathbf{x}; \hat{\boldsymbol{\eta}})\, d\mathbf{x} = 1,$$

where S_1 is the support of $\mathbf{X}$ given $Y = 1$, and the normalisation constraint arises since $r(\mathbf{x}; \hat{\boldsymbol{\eta}})\, f_0(\mathbf{x}; \hat{\boldsymbol{\eta}})$ is an estimate of $f_1(\mathbf{x})$. Provided common support $S_1 = S_0$, where S_0 is the support of $\mathbf{X}$ given $Y = 0$, one can use the empirical distribution function (EDF) of X over $\{\mathbf{x}_i : y_i = 1\}$ in place of f_1 for D, and that over $\{\mathbf{x}_i : y_i = 0\}$ in place of f_0 for the constraint. Having obtained $\hat{r}_{\mathbf{x}} = r(\mathbf{x}; \hat{\boldsymbol{\eta}})$, one can classify any unit given the associated feature vector $\mathbf{x}$ based on $\Pr(Y = 1 \mid \mathbf{x}; \hat{p}, \hat{r}_{\mathbf{x}})$, where $\hat{p}$ is an estimate of the prevalence $p = \Pr(Y = 1)$.

We describe how the idea of MEC for supervised learning can be adapted to record linkage problem in the following subsections.

3.1 Probability ratio for record linkage

For supervised learning based MEC to record linkage, suppose M is observed for the given Ω, and the trained classifier is to be applied to the record pairs outside of Ω. To fix the idea, suppose B is a non-probability sample that overlaps with the population P, and A is a probability sample from P with known inclusion probabilities. While $\boldsymbol{\gamma}_M = \{\boldsymbol{\gamma}_{ab} : (a, b) \in M\}$ may be considered as an IID sample, since each (a, b) in M refers to a distinct entity, this is not the case with $\{\boldsymbol{\gamma}_{ab} : (a, b) \notin M\}$, whose *joint* distribution is troublesome to model.

Probability ratio (I)

Let $r_q(\boldsymbol{\gamma})$ be the *probability ratio* given by

$$r_q(\boldsymbol{\gamma}) = \frac{m(\boldsymbol{\gamma})}{q(\boldsymbol{\gamma})},$$

where $m(\boldsymbol{\gamma})$ is the probability mass function of $\boldsymbol{\gamma}_{ab} = \boldsymbol{\gamma}$ given $g_{ab} = 1$, and $q(\boldsymbol{\gamma})$ is that over $\boldsymbol{\gamma}_\Omega = \{\boldsymbol{\gamma}_{ab} : (a, b) \in \Omega\}$. The KL divergence measure from $q(\boldsymbol{\gamma})$ to $m(\boldsymbol{\gamma})$ and the normalisation constraint are

$$D_f = \sum_{\gamma \in S(M)} m(\gamma) \log r_q(\gamma) \quad \text{and} \quad \sum_{\gamma \in S(M)} \hat{q}(\gamma)\, \hat{r}_q(\gamma) = 1,$$

where $S(M)$ is the support of γ_{ab} given $g_{ab} = 1$. This set-up allows $S(M)$ to be a subset of S, where S is the support of all possible γ_{ab}. It follows that, based on the IID sample γ_M of size $n_M = |M|$, the objective function to be *minimized* for r_q can be given by

$$Q_f = \sum_{(a,b) \in M} \frac{f(\gamma_{ab})}{n_M(\gamma_{ab})} r_q(\gamma_{ab}) - \frac{1}{n_M} \sum_{(a,b) \in M} \log r_q(\gamma_{ab}), \tag{3.1}$$

where $n_M(\boldsymbol{\gamma}_{ab}) = \sum_{(i,j) \in M} \mathbb{I}(\boldsymbol{\gamma}_{ij} = \boldsymbol{\gamma}_{ab})$ based on the observed support $S(M)$.

Probability ratio (II)

Provided $S(M) \subseteq S(U)$, where $S(U)$ is the support of γ_{ab} over U, one can let the probability ratio be given by

$$r(\gamma) = \frac{m(\gamma)}{u(\gamma)}$$

where $u(\gamma)$ is the probability of $\gamma_{ab} = \gamma$ given $g_{ab} = 0$. We have

$$r_q(\gamma) = \frac{m(\gamma)}{q(\gamma)} = \frac{m(\gamma)}{\pi m(\gamma) + (1-\pi) u(\gamma)} = \frac{r(\gamma)}{\pi(r(\gamma) - 1) + 1}$$

where $q(\gamma) = \pi m(\gamma) + (1-\pi) u(\gamma)$, so that $r_q(\gamma)$ and $r(\gamma)$ are one-to-one. Meanwhile, the KL divergence measure from $u(\gamma)$ to $m(\gamma)$ is given by

$$D = \sum_{\gamma \in S(M)} m(\gamma) \log r(\gamma)$$

and the objective function to be *minimized* for r can now be given by

$$Q = \sum_{(a,b) \in M} \frac{u(\gamma_{ab})}{n_M(\gamma_{ab})} r(\gamma_{ab}) - \frac{1}{n_M} \sum_{(a,b) \in M} \log r(\gamma_{ab}). \tag{3.2}$$

Model of γ: Under the multinomial model, one can simply use the EDF of γ over γ_Ω as $f(\gamma)$, for each distinct level of γ, as long as $|\Omega|$ is large compared to $|S|$. Similarly for $m(\gamma)$ over γ_M and $u(\gamma)$ over U. For linkage outside of Ω, the estimated $m(\gamma)$ from $M(\Omega)$ applies, if the selection of A from P is non-informative.

For γ made up of K binary agreement indicators, $\gamma_k = 0, 1$ for $k = 1, \ldots, K$, there are up to 2^K distinct levels of γ, which can sometimes be relatively large compared to $|M|$. A more parsimonious model of $m(\gamma; \boldsymbol{\theta})$ that is commonly used is given by

$$m(\gamma; \boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_k^{\gamma_k} (1 - \theta_k)^{1-\gamma_k} \tag{3.3}$$

where $\theta_k = \Pr\left(\gamma_{ab,k} = 1 \mid g_{ab} = 1\right)$, and $\gamma_{ab,k}$ is the kth component of $\boldsymbol{\gamma}_{ab}$. It is possible to model θ_k based on the distributions of the key variables that give rise to $\boldsymbol{\gamma}$, which makes use of the differential frequencies of their values, such as the fact that some names are more common than others. Similarly, $u(\boldsymbol{\gamma}; \boldsymbol{\xi})$ can be modeled as in (3.3) with parameters ξ_k instead of θ_k, where $\xi_k = \Pr\left(\gamma_{ab,k} = 1 \mid g_{ab} = 0\right)$.

Note that (3.3) implies conditional independence among agreement indicators. Winkler (1993) and Winkler (1994) demonstrated that even when the conditional independence assumption does not hold, results based on conditional independence assumption are quite robust. More complicated models that allow for correlated γ_k can also be considered. See Armstrong and Mayda (1993) and Larsen and Rubin (2001) for discussion of those models. See Xu, Li, Shen, Hui and Grannis (2019) for a recent study which compares models with or without correlated γ_k.

3.2 MEC sets for record linkage

Provided there are no duplicated records in either A or B, a *classification set* for record linkage, denoted by $\hat{M}$, consists of record pairs from Ω, where any record in A or B appears at most in one record pair in $\hat{M}$. Let the *entropy* of a classification set $\hat{M}$ be given by

$$D_{\hat{M}} = \frac{1}{\left|\hat{M}\right|} \sum_{(a,b)\in\hat{M}} \log r\left(\boldsymbol{\gamma}_{ab}\right). \tag{3.4}$$

A MEC set of given size $n^* = \left|\hat{M}\right|$ is the first classification set that is of size n^*, obtained by deduplication in the descending order of $r\left(\boldsymbol{\gamma}_{ab}\right)$ over Ω. It is possible to have $(a,b') \notin \hat{M}$ and $r\left(\boldsymbol{\gamma}_{ab'}\right) > r\left(\boldsymbol{\gamma}_{a',b'}\right)$ for $(a',b') \in \hat{M}$, if there exists $(a,b) \in \hat{M}$ with $r\left(\boldsymbol{\gamma}_{ab}\right) > r\left(\boldsymbol{\gamma}_{ab'}\right)$.

A MEC set of size n^* is not necessarily the largest possible classification set with the maximum entropy, to be referred to as a *maximal* MEC set, which is the largest classification set such that $r\left(\boldsymbol{\gamma}_{ab}\right) = \max_{\boldsymbol{\gamma}} r\left(\boldsymbol{\gamma}\right)$ for every (a,b) in it. In practice, a maximal MEC set is given by the first pass of *deterministic linkage*, which only consists of the record pairs with perfect *and* unique agreement of all the key variables.

Probabilistic linkage methods for MEC set are useful if one would like to allow for additional links, even though their key variables do not agree perfectly with each other. For the uncertainty associated with a given MEC set $\hat{M}$, we consider two types of errors. First, we define the *false link rate (FLR)* among the links in $\hat{M}$ to be

$$\psi = \frac{1}{\left|\hat{M}\right|} \sum_{(a,b)\in\hat{M}} \left(1 - g_{ab}\right) \tag{3.5}$$

which is different to μ by (2.1) where the denominator is $|U|$. Second, the *missing match rate (MMR)* of $\hat{M}$, which is related to the false non-link probability λ in (2.1), is given by

$$\tau = 1 - \frac{1}{n_M} \sum_{(a,b)\in\hat{M}} g_{ab}. \tag{3.6}$$

While μ and λ in (2.1) are theoretical probabilities, the FLR and MMR are actual errors.

It is instructive to consider the situation, where one is asked to form MEC sets in Ω given all the necessary estimates related to the probability ratio $r(\boldsymbol{\gamma})$, which can be obtained under the SL setting, without being given n_M, g_Ω or M directly.

First, the perfect MEC set should have the size n_M. Let $n(\boldsymbol{\gamma}) = \sum_{(a,b)\in\Omega} \mathbb{I}(\boldsymbol{\gamma}_{ab} = \boldsymbol{\gamma})$. One can obtain n_M as the solution to the following fixed-point equation:

$$n_M = \sum_{(a,b)\in\Omega} \hat{g}(\boldsymbol{\gamma}_{ab}) = \sum_{\gamma\in S} n(\boldsymbol{\gamma})\,\hat{g}(\boldsymbol{\gamma}) \tag{3.7}$$

where

$$\hat{g}(\boldsymbol{\gamma}) := \Pr\left(g_{ab} = 1 \,\middle|\, \boldsymbol{\gamma}_{ab} = \boldsymbol{\gamma}\right) = \frac{\pi r(\boldsymbol{\gamma})}{\pi\left(r(\boldsymbol{\gamma})-1\right)+1} = \frac{n_M r(\boldsymbol{\gamma})}{n_M\left(r(\boldsymbol{\gamma})-1\right)+n} \tag{3.8}$$

and the probability is defined with respect to completely random sampling of a single record pair from Ω. To see that $\hat{g}(\boldsymbol{\gamma})$ by (3.8) satisfies (3.7), notice $\hat{g}(\boldsymbol{\gamma}) = n_M m(\boldsymbol{\gamma}) / n(\boldsymbol{\gamma})$ satisfies (3.7) for any well defined $m(\boldsymbol{\gamma})$, and $n(\boldsymbol{\gamma})/n = \pi m(\boldsymbol{\gamma}) + (1-\pi)u(\boldsymbol{\gamma})$ by definition.

Next, apart from a maximal MEC set, one would need to accept discordant pairs. In the SL setting, one observes the EDF of $\boldsymbol{\gamma}$ over M, giving rise to $\hat{\theta}_k = n_M(1;k)/n_M$, where $n_M(1;k)$ is the number of agreements on the kth key variable over M. The perfect MEC set $\hat{M}$ should have these agreement rates. We have then, for $k = 1,\ldots,K$,

$$\hat{\theta}_k = \frac{1}{\left|\hat{M}\right|}\sum_{(a,b)\in\hat{M}} \mathbb{I}(\gamma_{ab,k} = 1) \quad \text{for} \quad \left|\hat{M}\right| = n_M. \tag{3.9}$$

Thus, no matter how one models $m(\boldsymbol{\gamma})$, the perfect MEC set should satisfy jointly the $K+1$ equations defined by (3.7) and (3.9), given the knowledge of $r(\boldsymbol{\gamma})$.

4. MEC for unsupervised record linkage

Let $\mathbf{z}$ be the K-vector of key variables, which may be imperfect for two reasons: it is not rich enough if the true $\mathbf{z}$-values are not unique for each distinct entity underlying the two files to be linked, or it may be subjected to errors if the observed $\mathbf{z}$ is not equal to its true value. Let A contain only the distinct $\mathbf{z}$-vectors from the first file, after removing any other record that has a duplicated $\mathbf{z}$-vector to some record that is retained in A. In other words, if the first file initially contains two or more records with exactly the same value of the combined key, then only one of them will be retained in A for record linkage to the second file. Similarly let B contain the unique records from the second file. The reason for *separate deduplication of keys* is that no comparisons between the two files can distinguish among the duplicated $\mathbf{z}$ in either file, which is an issue to be resolved otherwise.

Given A and B preprocessed as above, the maximal MEC set M_1 only consists of the record pairs with the perfect agreement of all the key variables. For probabilistic linkage beyond M_1, one can follow the same scheme of MEC in the supervised setting, as long as one is able to obtain an estimate of the probability ratio, given which one can form the MEC set of any chosen size. Nevertheless, to estimate the associated FLR (3.5) and MMR (3.6), an estimate of n_M is also needed.

4.1 Algorithm of unsupervised MEC

The idea now is to apply (3.7) and (3.9) jointly. Since setting $\hat{n}_M = |M_1|$ and $\hat{\theta}_k \equiv 1$ associated with the maximal MEC set satisfies (3.7) and (3.9) automatically, probabilistic linkage requires one to assume $n_M > |M_1|$ and $\theta_k < 1$ for at least some of $k = 1, \ldots, K$. Moreover, unless there is external information that dictates it otherwise, one can only assume common support $S(M) = S(U)$ in the unsupervised setting. Let

$$r(\boldsymbol{\gamma}) = m(\boldsymbol{\gamma}; \boldsymbol{\theta}) / u(\boldsymbol{\gamma}; \boldsymbol{\xi}) \tag{4.1}$$

where the probability of observing $\boldsymbol{\gamma}$ is $m(\boldsymbol{\gamma}; \boldsymbol{\theta})$ by (3.3) given that a randomly selected record pair from Ω belongs to M, and $u(\boldsymbol{\gamma}; \boldsymbol{\xi})$ otherwise, similarly given by (3.3) with parameters ξ_k instead of θ_k. An iterative algorithm of unsupervised MEC is given below.

I. Set $\boldsymbol{\theta}^{(0)} = \left(\theta_1^{(0)}, \ldots, \theta_K^{(0)}\right)$ and $n_M^{(0)} = |M_1|$, where M_1 is the maximal MEC set.

II. For the tth iteration, let $g_{ab}^{(t)} = 1$ if $(a, b) \in M^{(t)}$, and 0 otherwise.

   i. Update $u\left(\boldsymbol{\gamma}; \boldsymbol{\xi}^{(t)}\right)$ by using (4.4), which is discussed below, given $\mathbf{g}^{(t)} = \left\{g_{ab}^{(t)} : (a, b) \in \Omega\right\}$, and calculate

$$\theta_k^{(t)} = \frac{1}{\left|M^{(t)}\right|} \sum_{(a,b) \in \Omega} g_{ab}^{(t)} \, \mathbb{I}\left(\gamma_{ab,k} = 1\right), \tag{4.2}$$

   which maximize D_M in (3.4) for given $u\left(\boldsymbol{\gamma}; \boldsymbol{\xi}^{(t)}\right), M^{(t)} = \left\{(a, b) \in \Omega : g_{ab}^{(t)} = 1\right\}$ and $\left|M^{(t)}\right| = \sum_{(a,b) \in \Omega} g_{ab}^{(t)}$. Once $\boldsymbol{\theta}^{(t)}$ and $\boldsymbol{\xi}^{(t)}$ are obtained, we can update $n_M^{(t)} = \sum_{\boldsymbol{\gamma}} n(\boldsymbol{\gamma}) \, \hat{g}^{(t)}(\boldsymbol{\gamma})$, where

$$\hat{g}^{(t)}(\boldsymbol{\gamma}) \equiv \hat{g}\left(\boldsymbol{\gamma}; \boldsymbol{\theta}^{(t)}, \boldsymbol{\xi}^{(t)}\right) = \min\left\{\frac{\left|M^{(t)}\right| r^{(t)}(\boldsymbol{\gamma})}{\left|M^{(t)}\right|\left(r^{(t)}(\boldsymbol{\gamma}) - 1\right) + n}, 1\right\}$$

$$r^{(t)}(\boldsymbol{\gamma}) \equiv r\left(\boldsymbol{\gamma}; \boldsymbol{\theta}^{(t)}, \boldsymbol{\xi}^{(t)}\right) = \frac{m\left(\boldsymbol{\gamma}; \boldsymbol{\theta}^{(t)}\right)}{u\left(\boldsymbol{\gamma}; \boldsymbol{\xi}^{(t)}\right)}.$$

   ii. For given $\boldsymbol{\theta}^{(t)}, \boldsymbol{\xi}^{(t)}$ and $n_M^{(t)}$, we find the MEC set $M^{(t+1)} = \left\{(a, b) \in \Omega : g_{ab}^{(t+1)} = 1\right\}$ such that $\left|M^{(t+1)}\right| = n_M^{(t)}$ by deduplication in the descending order of $r^{(t)}(\boldsymbol{\gamma}_{ab})$ over Ω. It maximizes the entropy denoted by $Q^{(t)}(\mathbf{g})$:

$$Q^{(t)}(\mathbf{g}) \equiv Q\left(\mathbf{g} \mid \boldsymbol{\psi}^{(t)}\right) = \frac{1}{n_M^{(t)}} \sum_{(a,b)\in\Omega} g_{ab} \log r^{(t)}(\boldsymbol{\gamma}_{ab}), \tag{4.3}$$

with respect to $\mathbf{g}$.

III. Iterate until $n_M^{(t)} = n_M^{(t+1)}$ or $\left\| \boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(t+1)} \right\| < \epsilon$, where ϵ is a small positive value.

A theoretical convergence property of the proposed algorithm and its proof are presented in the supplementary materials.

Notice that, insofar as $\Omega = M \cup U$ is highly imbalanced, where the prevalence of $g_{ab} = 1$ is very close to 0, one could simply ignore the contributions from M and use

$$\hat{\xi}_k = \frac{1}{n} \sum_{(a,b)\in\Omega} \mathbb{I}\left(\gamma_{(ab,k)} = 1\right) \tag{4.4}$$

under the model (3.3) of $u(\boldsymbol{\gamma};\boldsymbol{\xi})$, in which case there is no updating of $u\left(\boldsymbol{\gamma};\boldsymbol{\xi}^{(t)}\right)$. Other possibilities of estimating $u(\boldsymbol{\gamma};\boldsymbol{\xi})$ will be discussed in Section 5.2.

Table 4.1 provides an overview of MEC for record linkage in the supervised or unsupervised setting. In the supervised setting, one observes $\boldsymbol{\gamma}$ for the matched record pairs in M, so that the probability $m(\boldsymbol{\gamma})$ can be estimated from them directly. Whereas, for MEC in the unsupervised setting, one cannot separate the estimation of $m(\boldsymbol{\gamma})$ and n_M.

Table 4.1
MEC for record linkage in supervised or unsupervised setting

	Supervised	**Unsupervised**
$\boldsymbol{\Omega} = \boldsymbol{M} \cup \boldsymbol{U}$	**Observed**	**Unobserved**
Probability ratio	$r_q(\boldsymbol{\gamma})$ generally applicable $r(\boldsymbol{\gamma})$ given $S(M) \subseteq S(U)$	$r(\boldsymbol{\gamma})$ generally assuming $S(M) = S(U)$
Model of $\boldsymbol{\gamma}$	Multinomial if only discrete comparison scores Directly or via key variables and measurement errors	
MEC set	Guided by FLR and MMR Require estimate of n_M in addition	
Estimation	$m(\boldsymbol{\gamma};\boldsymbol{\theta})$ from $\boldsymbol{\gamma}_M$ in Ω n_M by (3.7) outside Ω	$m(\boldsymbol{\gamma};\boldsymbol{\theta})$ and n_M jointly by (3.7) and (3.9)

4.2 Error rates

MEC for record linkage should generally be guided by the error rates, FLR and MMR, without being restricted to the estimate of n_M.

Note that $\left\{\hat{g}_{ab} : (a,b) \in \hat{M}\right\}$ of any MEC set $\hat{M}$ are among the largest ones over Ω, because MEC follows the descending order of $\hat{r}_{ab}$, except for necessary deduplication when there are multiple pairs involving a given record. To exercise greater control of the FLR, let ψ be the target FLR, and consider the following bisection procedure.

i. Choose a threshold value c_ψ and form the corresponding MEC set $\hat{M}(c_\psi)$, where $\hat{r}_{ab} \geq c_\psi$ for any $(a,b) \in \hat{M}(c_\psi)$.

ii. Calculate the estimated FLR of the resulting MEC set $\hat{M}$ as

$$\hat{\psi} = \frac{1}{|\hat{M}|} \sum_{(a,b)\in\hat{M}} (1 - \hat{g}_{ab}). \tag{4.5}$$

If $\hat{\psi} > \psi$, then increase c_ψ; if $\hat{\psi} < \psi$, then reduce c_ψ.

Iteration between the two steps would eventually lead to a value of c_ψ that makes $\hat{\psi}$ as close as possible to ψ, for the given probability ratio $\hat{r}(\boldsymbol{\gamma})$.

The final MEC set $\hat{M}$ can be chosen in light of the corresponding FLR estimate $\hat{\psi}$. It is also possible to take into consideration the estimated MMR given by

$$\hat{\tau} = 1 - \sum_{(a,b)\in\hat{M}} \hat{g}_{ab} / \hat{n}_M \tag{4.6}$$

where $\hat{n}_M$ is given by unsupervised MEC algorithm. Note that if $|\hat{M}| = \hat{n}_M$, then we shall have $\hat{\psi} = \hat{\tau}$; but not if $\hat{M}$ is guided by a given target value of FLR or MMR.

In Section 6.2, we investigate the performance of the MEC sets guided by the error rates through simulations.

5. Discussion

Below we discuss and compare two other approaches in the unsupervised setting, including the ways by which some of their elements can be incorporated into the MEC approach. Other less practical approaches are discussed in the supplementary material.

5.1 The classical approach

Recall Problems I and II of the classical approach mentioned in Section 2.

From a practical point of view, Problem I can be dealt with by any deduplication method of the set M^* of classified records pairs, where $\hat{r}(\boldsymbol{\gamma}_{ab})$ is above a threshold value for all $(a,b) \in M^*$. As "an advance over previous ad hoc assignment methods", Jaro (1989) chooses the linked set $\hat{M}^* \subseteq M^*$, which maximises the sum of $\log \hat{r}(\boldsymbol{\gamma}_{ab})$ subject to the constraint of one-one link. Since $\hat{g}_{ab}$ is a monotonic function of $\hat{r}(\boldsymbol{\gamma}_{ab})$, this amounts to choose $\hat{M}^*$ which maximises the expected number of matches in it, denoted by

$$n_M^* = \sum_{(a,b)\in\hat{M}^*} \hat{g}_{ab}$$

But n_M^* is still not connected to the probabilities of false links and non-links defined by (2.1). As illustrated below, neither does it directly control the errors of the linked $\hat{M}^*$.

Consider linking two files with 100 records each. Suppose Jaro's assignment method yields $\left|\hat{M}^*\right| = 100$ on one occasion, where 80 links have $\hat{g}_{ab} \approx 1$ and 20 links have $\hat{g}_{ab} \approx 0.75$, such that $n_M^* \approx 95$. Suppose it yields 90 links with $\hat{g}_{ab} \approx 1$ and 10 links with $\hat{g}_{ab} \approx 0.5$ on another occasion, where $n_M^* \approx 95$. Clearly, n_M^* does not directly control the linkage errors in $\hat{M}^*$. Moreover, there is no compelling reason to accept 100 links on both these occasions, simply because 100 one-one links are possible.

In forming the MEC set one deals with Problem I directly, based on the concept of maximum entropy that has relevance in many areas of scientific investigation. The implementation is simple and fast for large datasets. The estimated error rates FLR (4.5) and MMR in (4.6) are directly defined for a given MEC set.

Problem II concerns the parameter estimation. As explained earlier, applying the EM algorithm based on the objective function (2.2) proposed by Winkler (1988) and Jaro (1989) is *not* a valid approach of maximum likelihood estimation (MLE). One may easily compare this WJ-procedure to that given in Section 4.1, where both adopt the same model (3.3) and the same estimator of $u(\boldsymbol{\gamma};\boldsymbol{\xi})$ via $\hat{\xi}_k$ given by (4.4). It is then clear that the same formula is used for updating $n_M^{(t)}$ at each iteration, but a different formula is used for

$$\theta_k^{(t)} = \frac{1}{n_M^{(t)}} \sum_{(a,b)\in\Omega} \hat{g}_{ab}^{(t)} \gamma_{ab,k} \tag{5.1}$$

where the numerator is derived from *all* the pairs in Ω, whereas $\theta_k^{(t)}$ given by (4.2) uses only the pairs in the MEC set $M^{(t)}$. Notice that the two differ only in the unsupervised setting, but they would become the same in the supervised setting, where one can use the observed binary g_{ab} instead of the estimated fractional $\hat{g}_{ab}$.

Thus, one may incorporate the WJ-procedure as a variation of the unsupervised MEC algorithm, where the formulae (5.1) and (4.4) are chosen specifically. This is the reason why it can give reasonable parameter estimates in many situations, despite its misconception as the MLE. Simulations will be used later to compare empirically the two formulae (4.2) and (5.1) for $\theta_k^{(t)}$.

5.2 An approach of MLE

Below we derive another estimator of ξ_k by the ML approach, which can be incorporated into the proposed MEC algorithm, instead of (4.4). This requires a model of the key variables, which explicates the assumptions of key-variable errors. Let z_k be the kth key variable which takes value $1,\ldots,D_k$. Copas and Hilton (1990) envisage a non-informative hit-miss generation process, where the observed z_k can take the true value despite the perturbation. Copas and Hilton (1990) demonstrate that the hit-miss model is plausible in the SL (Supervised Learning) setting based on labelled datasets.

We adapt the hit-miss model to the unsupervised setting as follows. First, for any $(a,b)\in M$, let $\alpha_k = \Pr\left(e_{ab,k}=1\right)$, where $e_{ab,k}=1$ if the associated pair of key variables are subjected to *any form of*

perturbation that could potentially cause disagreement of the kth key variable, and $e_{ab,k} = 0$ otherwise. Let

$$\theta_k = (1-\alpha_k) + \alpha_k \sum_{d=1}^{D_k} m_{kd}^2 = 1 - \alpha_k \left(1 - \sum_{d=1}^{D_k} m_{kd}^2\right)$$

where we assume that α_k must be positive for some $k = 1, \ldots, K,$ and

$$m_{kd} = \Pr\left(z_{ik} = d \mid g_{ab} = 1, e_{ab,k} = 1\right) = \Pr\left(z_{ik} = d \mid g_{ab} = 1, e_{ab,k} = 0\right)$$

for $i = a$ or b. Next, for any record i in either A or B, let $\delta_i = 1$ if it has a match in the other file and $\delta_i = 0$ otherwise. Given $\delta_i = 0,$ with or without perturbation, let $\Pr\left(z_{ik} = d \mid \delta_i = 0\right) = u_{kd}$. We have $\beta_{kd} := m_{kd} \equiv u_{kd}$ if δ_i is *non-informative*. A slightly more relaxed assumption is that δ_i is only non-informative in one of the two files. To be more resilient against its potential failure, one can assume m_{kd} to hold for all the records in the *smaller* file, and allow u_{kd} to differ for the records with $\delta_i = 0$ in the *larger* file. Suppose $n_A < n_B$. Let

$$p = \Pr\left(\delta_b = 1\right) = E\left(n_M\right) / n_B = n_A \pi$$

be the probability that a record in B has a match in A. One may assume $\mathbf{z}_A = \{\mathbf{z}_a : a \in A\}$ to be independent over A, giving

$$\ell_A = \sum_{a \in A} \sum_{k=1}^{K} \log m_{ak}$$

where $m_{ak} = \sum_{d=1}^{D_k} m_{kd}\ \mathbb{I}(z_{ak} = d)$. The complete-data log-likelihood based on $\left(\delta_B, \mathbf{z}_B\right)$ is

$$\ell_B = \sum_{b \in B} \delta_b \log\left(p \prod_{k=1}^{K} m_{bk}\right) + \sum_{b \in B} \left(1 - \delta_b\right) \log\left((1-p) \prod_{k=1}^{K} u_{bk}\right) \tag{5.2}$$

where $m_{bk} = \sum_{d=1}^{D_k} m_{kd}\ \mathbb{I}(z_{bk} = d)$ and $u_{bk} = \sum_{d=1}^{D_k} u_{kd}\ \mathbb{I}(z_{bk} = d)$, based on an assumption of independent $\left(\delta_b, \mathbf{z}_b\right)$ across the entities in B.

Under separate modelling of $\mathbf{z}_A$ and $\left(\mathbf{z}_B, \delta_B\right)$, let $\hat{m}_{kd}$ be the MLE based on ℓ_A, given which an EM-algorithm for estimating p and u_{kd} follows from (5.2) by treating δ_B as the missing data. However, the estimation is feasible only if $\{u_{kd}\}$ and $\{m_{kd}\}$ are not exactly the same; whereas the MLE of n_M has a large variance, when $\{m_{kd}\}$ and $\{u_{kd}\}$ are close to each other, even if they are not exactly equal.

Meanwhile, the closeness between $\{m_{kd}\}$ and $\{u_{kd}\}$ does not affect the MEC approach, where $\hat{n}_M$ is obtained from solving (3.7) given $\hat{r}(\boldsymbol{\gamma}) = \hat{m}(\boldsymbol{\gamma}) / \hat{u}(\boldsymbol{\gamma})$, where $\hat{u}(\boldsymbol{\gamma})$ is indeed most reliably estimated when $\{m_{kd}\} = \{u_{kd}\}$. Moreover, one can incorporate a *profile EM-algorithm*, based on (5.2) given $n_M^{(t)}$, to update $u\left(\boldsymbol{\gamma}; \boldsymbol{\xi}^{(t)}\right)$ in the unsupervised MEC algorithm of Section 4.1. At the tth iteration, where $t \geq 1$, given $p^{(t)} = n_M^{(t)} / \max\left(n_A, n_B\right)$ and $\hat{m}_{kd}$ estimated from the smaller file A, obtain $u_{kd}^{(t)}$ by

$$\xi_k^{(t)} = \left(\left(1 - p^{(t)}\right) \sum_{d=1}^{D_k} u_{kd}^{(t)} \hat{m}_{kd} + p^{(t)} \left(1 - \frac{1}{n_A}\right) \sum_{d=1}^{D_k} \hat{m}_{kd}^2\right) \Big/ \left(1 - p^{(t)} / n_A\right). \tag{5.3}$$

6. Simulation study

6.1 Set-up

To explore the practical feasibility of the unsupervised MEC algorithm for record linkage, we conduct a simulation study based on the data sets listed in Table 6.1, which are disseminated by ESSnet-DI (McLeod, Heasman and Forbes, 2011) and freely available online. Each record in a data set has associated synthetic key variables, which may be distorted by missing values and typos when they are created, in ways that imitate real-life errors (McLeod et al., 2011).

Table 6.1
Data set description (size in parentheses)

Data set	Description
Census (25,343)	A fictional data set to represent some observations from a decennial Census.
CIS (24,613)	Fictional observations from Customer Information System, combined administrative data from the tax and benefit systems.
PRD (24,750)	Fictional observations from Patient Register Data of the National Health Service.

We consider the linkage keys forename, surname, sex, and date of birth (DOB). To model the key variables, we divide DOB into 3 key variables (Day, Month, Year). For text variables such as forename and surname, we divide them into 4 key variables by using the Soundex coding algorithm (Copas and Hilton, 1990, page 290), which reduces a name to a code consisting of the leading letter followed by three digits, e.g. Copas ≡ C120, Hilton ≡ H435. The twelve key variables for record linkage are presented in Table 6.2.

Table 6.2
Twelve key variables available in the three data sets

Variable		Description	No. of Categories
PERNAME1	1	First letter of forename	26
	2	First digit of Soundex code of forename	7
	3	Second digit of Soundex code of forename	7
	4	Third digit of Soundex code of forename	7
PERNAME2	1	First letter of surname	26
	2	First digit of Soundex code of surname	7
	3	Second digit of Soundex code of surname	7
	4	Third digit of Soundex code of surname	7
SEX		Male/Female	2
DOB	DAY	Day of birth	31
	MON	Month of birth	12
	YEAR	Year of birth (1910 ~ 2012)	103

We set up two scenarios to generate linkage files. We use the unique identification variable (PERSON-ID) for sampling, which are available in all the three data sets. We sample $n_A = 500$ and $n_B = 1{,}000$ individuals from PRD and CIS, respectively. Let p_A be the proportion of records in the smaller file (PRD) that are also selected in the larger file (CIS), by which we can vary the degree of overlap, i.e. the set of matched individuals AB, between A and B. We use $p_A = 0.8, 0.5$ or 0.3 under either scenario.

Scenario-I (Non-informative)

- Sample $n_0 = n_B / p_A$ individuals randomly from Census.
- Sample n_A randomly from these n_0 as the individuals of PRD, denoted by A.
- Sample n_B randomly from these n_0 as the individuals of CIS, denoted by B.

Under this scenario both δ_a and δ_b are non-informative for the key-variable distribution. For any given p_A, we have $E(n_M) = n_A p_A$ and $\pi = E(n_M)/n_0$, where n_M is the random number of matched individuals between the simulated files A and B.

Scenario-II (Informative)

- Sample n_A randomly from Census $\cap$ PRD $\cap$ CIS, denoted by A from PRD.
- Sample $n_M = n_A p_A$ randomly from A as the matched individuals, denoted by AB.
- Sample $n_B - n_M$ randomly from CIS $\setminus$ A having SEX $= F$, YEAR ≤ 1970, and odd MON, denoted by B_0. Let $B = AB \cup B_0$ be the sampled individuals of CIS.

Under this scenario the key-variable distribution is the same in A, whether or not $\delta_a = 1$, but it is different for the records $b \in B_0$, or $\delta_b = 0$. Hence, scenario-II is informative. For any given p_A, we have fixed $n_M = n_A p_A$ and $\pi = p_A / n_B$.

6.2 Results: Estimation

For the unsupervised MEC algorithm given in Section 4.1, one can adopt (4.2) or (5.1) for updating $\theta_k^{(t)}$. Moreover, one can use (4.4) for $\hat{\xi}_k$ directly, or (5.3) for updating $\xi_k^{(t)}$ iteratively. In particular, choosing (5.1) and (4.4) effectively incorporates the procedure of Winkler (1988) and Jaro (1989) for parameter estimation. Note that the MEC approach still differs to that of Jaro (1989), with respect to the formation of the linked set $\hat{M}$.

Table 6.3 compares the performance of the unsupervised MEC algorithm, using different formulae for $\theta_k^{(t)}$ and $\xi_k^{(t)}$, where the size of $\hat{M}$ is equal to the corresponding estimate $\hat{n}_M$. In addition, we include $\hat{\theta}_k = n_M(1;k)/n_M$ estimated directly from the matched pairs in M, as if M were available for supervised learning, together with (4.4) for $\hat{\xi}_k$. The true parameters and error rates are given in addition to their estimates.

Table 6.3
Parameters and averages of their estimates, averages of error rates and their estimates, over 200 simulations. Median of estimate of n_M given as $\tilde{n}_M$

Scenario I											Scenario II										
Parameter		Formulae		Estimation							Parameter		Formulae		Estimation						
π	$E(n_M)$	$\theta_k^{(t)}$	$\xi_k^{(t)}$	$\hat{\pi}$	$\hat{n}_M$	$\tilde{n}_M$	FLR	MMR	$\widehat{\text{FLR}}$	$\widehat{\text{MMR}}$	π	n_M	$\theta_k^{(t)}$	$\xi_k^{(t)}$	$\hat{\pi}$	$\hat{n}_M$	$\tilde{n}_M$	FLR	MMR	$\widehat{\text{FLR}}$	$\widehat{\text{MMR}}$
0.0008	400	$\hat{\theta}_k$	(4.4)	0.00080	400.0	397	0.0264	0.0266	0.0357	0.0357	0.0008	400	$\hat{\theta}_k$	(4.4)	0.00080	398.3	400	0.0230	0.0273	0.0326	0.0326
		(4.2)	(5.3)	0.00082	407.9	405	0.0425	0.0257	0.0509	0.0509			(4.2)	(5.3)	0.00080	401.4	401	0.0305	0.0277	0.0403	0.0403
		(4.2)	(4.4)	0.00083	414.7	407	0.0549	0.0244	0.0620	0.0620			(4.2)	(4.4)	0.00081	405.2	404	0.0379	0.0262	0.0467	0.0467
		(5.1)	(4.4)	0.00081	406.0	405	0.0399	0.0269	0.0503	0.0503			(5.1)	(4.4)	0.00080	401.4	401	0.0316	0.0286	0.0438	0.0438
0.0005	250	$\hat{\theta}_k$	(4.4)	0.00050	251.6	249	0.0340	0.0301	0.0370	0.0370	0.0005	250	$\hat{\theta}_k$	(4.4)	0.00050	249.6	250	0.0284	0.0302	0.0334	0.0334
		(4.2)	(5.3)	0.00052	258.3	255	0.0559	0.0296	0.0533	0.0533			(4.2)	(5.3)	0.00050	251.8	251	0.0383	0.0320	0.0410	0.0410
		(4.2)	(4.4)	0.00053	266.9	256.5	0.0742	0.0277	0.0680	0.0680			(4.2)	(4.4)	0.00052	257.7	253	0.0513	0.0295	0.0516	0.0516
		(5.1)	(4.4)	0.00052	261.7	259	0.0676	0.0305	0.0636	0.0636			(5.1)	(4.4)	0.00051	255.4	253.5	0.0510	0.0336	0.0520	0.0520
0.0003	150	$\hat{\theta}_k$	(4.4)	0.00030	152.3	151	0.0439	0.0356	0.0381	0.0381	0.0003	150	$\hat{\theta}_k$	(4.4)	0.00030	150.5	150	0.0382	0.0355	0.0350	0.0350
		(4.2)	(5.3)	0.00033	165.9	156.5	0.0873	0.0244	0.0620	0.0620			(4.2)	(5.3)	0.00031	153.0	153	0.0559	0.0377	0.0452	0.0452
		(4.2)	(4.4)	0.00041	205.4	161	0.1632	0.0308	0.1251	0.1251			(4.2)	(4.4)	0.00032	158.5	155	0.0708	0.0342	0.0558	0.0558
		(5.1)	(4.4)	0.00054	271.4	169	0.3015	0.0785	0.1639	0.1639			(5.1)	(4.4)	0.00038	189.3	156	0.1414	0.0524	0.0903	0.0903

As expected, the best results are obtained when the parameter θ_k is estimated directly from the matched pairs in M, i.e., $\hat{\theta}_k = n_M(1;k)/n_M$, together with (4.4) for $\hat{\xi}_k$, despite $\hat{\xi}_k$ by (4.4) is not exactly unbiased. Nevertheless, the approximate estimator $\hat{\xi}_k$ can be improved, since the profile-EM estimator given by (5.3) is seen to perform better across all the set-ups, where both are combined with (4.2) for $\theta_k^{(t)}$. When it comes to the two formulae of $\theta_k^{(t)}$ by (4.2) and (5.1), and the resulting n_M-estimators and the error rates FLR and MMR, we notice the followings.

- Scenario-I: When the size of the matched set M is relatively large at $p_A = 0.8$, there are only small differences in terms of the average and median of the two estimators of n_M, and the difference is just a couple of false links in terms of the linkage errors. Figures 6.1 shows that (4.2) results in a few larger errors of $\hat{n}_M$ than (5.1) over the 200 simulations, when $p_A = 0.8$ or $\pi = 0.0008$. As the size of the matched set M decreases, the averages and medians of the estimators of n_M resulting from (4.2) and (5.3) are closer to the true values than those of the other estimators. Especially when the matched set M is relatively small, where $\pi = 0.0003$, the formula (5.1) results in considerably worse estimation of n_M in every respect. While this is partly due to the use of (4.4) instead of (5.3), most of the difference is down to the choice of $\theta_k^{(t)}$, which can be seen from intermediary comparisons to the results based on (4.2) and (4.4).
- Scenario-II: The use of (4.2) and (5.3) for the unsupervised MEC algorithm performs better than using the other formulae in terms of both estimation of n_M and error rates across the three sizes of the matched set (Figure 6.2). Relatively greater improvement is achieved by using (4.2) and (5.3) for the smaller matched sets.

The results suggest that the unsupervised MEC algorithm tends to be more affected by the size of the matched set under Scenario-I than Scenario-II. Choosing (4.2) and (5.3), however, seems to yield the most

robust estimation of n_M and error rates against the small size of the matched set M, regardless the informativeness of key-variable errors. The reason must be the fact that the numerator of $\theta_k^{(t)}$ is calculated in (5.1) over all the pairs in Ω instead of the MEC set $M^{(t)}$, which seems more sensitive when the imbalance between M and U is aggravated, while the sizes of A and B remain fixed.

Figure 6.1 Box plots of $\hat{n}_M - n_M$ based on 200 Monte Carlo samples under Scenario I.

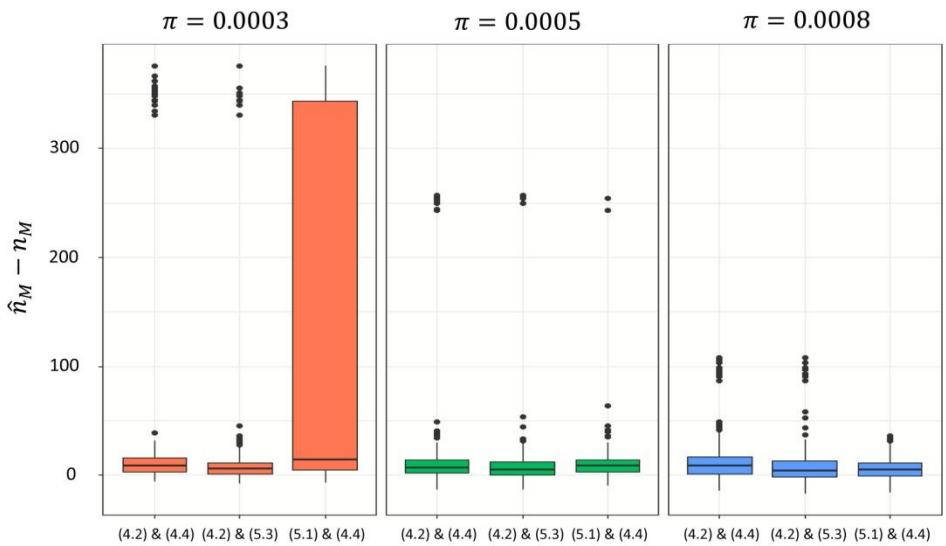
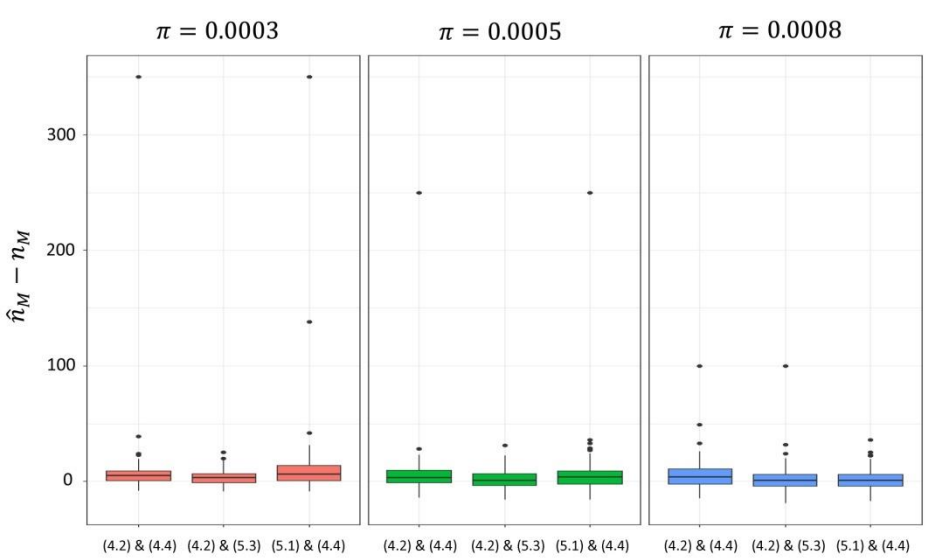


Figure 6.2 Box plots of $\hat{n}_M - n_M$ based on 200 Monte Carlo samples under Scenario II.



We also include the additional results obtained for $p_A = 0.2, 0.15,$ and 0.1 in the supplementary material. The estimate $\hat{n}_M$ (or $\hat{\pi}$) gets worse as p_A (or π) reduces, which is consistent with the previous findings of others, for example, Enamorado, Fifield and Imai (2019) showed that a greater degree of overlap between data sets leads to better merging results in terms of the error rates as well as the accuracy of their estimates. The problem is also highlighted by Sadinle (2017). Record linkage in cases of extremely low prevalence of true matches is a problem that needs to be studied more carefully on its own.

6.3 Results: MEC set

Aiming the MEC set $\hat{M}$ at the estimated size $\hat{n}_M$ is generally not a reasonable approach to record linkage. Record linkage should be guided directly by the associated uncertainty, i.e. the error rates FLR and MMR, based on their estimates (4.5) and (4.6), as described in Section 4.2. Note that this does require the estimation of n_M in addition to $r(\gamma)$.

We have $\widehat{\text{FLR}} = \widehat{\text{MMR}}$ in Table 6.3, because $|\hat{M}| = \hat{n}_M$ here. It can be seen that these follow the true FLR more closely than the MMR, especially when $\hat{n}_M$ is estimated using the formulae (4.2) and (5.3). This is hardly surprising. Take e.g. the maximal MEC set M_1 that consists of the pairs whose key variables agree completely and uniquely. Provided reasonably rich key variables, as the setting here, one can expect the FLR of M_1 to be low, such that even a naïve estimate $\widehat{\text{FLR}} = 0$ probably does not err much. Meanwhile, the true MMR has a much wider range from one application to another, because the difference between n_M and $|M_1|$ is determined by the extent of key-variable errors, such that the estimate of MMR depends more critically on that of n_M. The situation is similar for any MEC set beyond M_1, as long as $\hat{g}_{ab}$ remains very high for any $(a, b) \in \hat{M}$.

Table 6.4 shows the performance of the MEC set using the bisection procedure described in Section 4.2, across the same set-ups as in Table 6.3. We use only (4.2) for $\theta_k^{(t)}$ and (5.3) for $\xi_k^{(t)}$ to obtain the corresponding $\hat{n}_M$. We let the target FLR be $\psi = 0.05$ or 0.03, where the latter is clearly lower than the true FLR of $\hat{M}$ that is of the size $\hat{n}_M$ (Table 6.3), especially when the prevalence is relatively low (at $\pi = 0.0003$) under either scenario. The resulting true (FLR, MMR) and their estimates are given in Table 6.4.

Table 6.4
Parameters and averages of their estimates, averages of error rates and their estimates, over 200 simulations, $n = |\Omega| = n_A n_B$

Scenario I										Scenario II									
Parameter		Target	Estimation							Parameter		Target	Estimation						
π	$E(n_M)$	FLR	$\hat{n}_M$	$\lvert\hat{M}\rvert/n$	$\lvert\hat{M}\rvert$	FLR	MMR	$\widehat{\text{FLR}}$	$\widehat{\text{MMR}}$	π	n_M	FLR	$\hat{n}_M$	$\lvert\hat{M}\rvert/n$	$\lvert\hat{M}\rvert$	FLR	MMR	$\widehat{\text{FLR}}$	$\widehat{\text{MMR}}$
0.0008	400	0.05	407.9	0.00080	401.9	0.0313	0.0280	0.0393	0.0527	0.0008	400	0.05	401.4	0.00080	397.8	0.0239	0.0294	0.0337	0.0418
		0.03		0.00079	395.0	0.0196	0.0328	0.0271	0.0568			0.03		0.00079	393.1	0.0164	0.0334	0.0256	0.0451
0.0005	250	0.05	258.3	0.00050	251.9	0.0396	0.0326	0.0385	0.0576	0.0005	250	0.05	251.8	0.00050	248.6	0.0305	0.0361	0.0328	0.0447
		0.03		0.00049	246.7	0.0246	0.0374	0.0264	0.0650			0.03		0.00049	245.2	0.0226	0.0416	0.0245	0.0497
0.0003	150	0.05	165.9	0.00031	153.4	0.0533	0.0403	0.0389	0.0783	0.0003	150	0.05	153.0	0.00030	150.1	0.0445	0.0443	0.0333	0.0514
		0.03		0.00030	149.3	0.0355	0.0483	0.0256	0.0905			0.03		0.00029	147.4	0.0322	0.0489	0.0238	0.0588

It can be seen that the MEC algorithm guided by the FLR yields the MEC set $\hat{M}$, whose size $\left|\hat{M}\right|$ is close to the true n_M across all the set-ups. Indeed, under Scenario-I, the mean of $\left|\hat{M}\right|$ is closer to n_M than the mean (or median) of $\hat{n}_M$ over all the simulations, which results directly from parameter estimation, especially when the match set is relatively small (at $\pi = 0.0003$) and the performance of $\hat{n}_M$ is most sensitive. In other words, the fact that $\left|\hat{M}\right|$ differs to the estimate $\hat{n}_M$ is not necessarily a cause of concern for the MEC algorithm guided by targeting the FLR.

To estimate the MMR by (4.6), one can either use $\left|\hat{M}\right|$ as the estimate of n_M, or one can use $\hat{n}_M$ from parameter estimation based on (4.2) and (5.3). In the former case, one would obtain $\widehat{\mathrm{MMR}} = \widehat{\mathrm{FLR}}$. While this $\widehat{\mathrm{MMR}}$ is not unreasonable in absolute terms since $\left|\hat{M}\right|$ is close to n_M here, as can be seen from comparing the mean of $\widehat{\mathrm{FLR}}$ with that of the true MMR in Table 6.4, it has a drawback *a priori*, in that it decreases as the target FLR decreases, although one is likely to miss out on more true matches when more links are excluded from the MEC set $\hat{M}$. Using $\hat{n}_M$ from parameter estimation directly makes sense in this respect, since the true n_M must remain the same, regardless the target FLR. However, the estimator $\widehat{\mathrm{MMR}}$ could then become less reliable given relatively low prevalence π, where $\hat{n}_M$ could be sensitive in such situations.

In short, the estimation of FLR tends to be more reliable than that of MMR, especially if the prevalence π is relatively low in its theoretical range $0 < \pi \le \min\left(n_A, n_B\right)/n$. The following recommendations for unsupervised record linkage seem warranted.

- When forming the MEC set $\hat{M}$ according to the uncertainty of linkage, it is more robust to rely on the FLR, estimated by (4.5).
- The estimate of MMR given by (4.6), derived from the parameter estimate $\hat{n}_M$ based on (4.2) and (5.3) provides an additional uncertainty measure. However, one should be aware that this measure can be sensitive when the prevalence π is relatively low.
- Between two target values of the FLR, $\psi < \psi'$, more attention can be given to the estimate of additional missing matches in $\hat{M}(\psi)$ compared to $\hat{M}(\psi')$, given by

$$\sum_{(a,b)\in\hat{M}(\psi')} \hat{g}_{ab} - \sum_{(a,b)\in\hat{M}(\psi)} \hat{g}_{ab} = \sum_{(a,b)\in\hat{M}(\psi')\setminus\hat{M}(\psi)} \hat{g}_{ab}\,.$$

7. Final remarks

We have developed an approach of maximum entropy classification to record linkage. This provides a unified probabilistic record linkage framework both in the supervised and unsupervised settings, where a coherent classification set of links are chosen explicitly with respect to the associated uncertainty. The theoretical formulation overcomes some persistent flaws of the classical approaches. Furthermore, the proposed MEC algorithm is fully automatic, unlike the classical approach that generally requires clerical review to resolve the undecided cases.

An important issue that is worth further research concerns the estimation of relevant parameters in the model of key-variable errors that cause problems for record linkage. First, as pointed out earlier, treating record linkage as a classification problem allows one to explore many modern machine learning techniques. A key challenge in this respect is the fact that the different record pairs are not distinct "units", such that any powerful supervised learning technique needs to be adapted to the unsupervised setting, where it is impossible to estimate the relevant parameters based on the true matches and non-matches, including the number of matched entities. Next, the model of the key-variable errors or the comparison scores can be refined. Once these issues are resolved together, further improvements on the parameter estimation can hopefully be made, which will benefit both the classification of the set of links and the assessment of the associated uncertainty.

Another issue that is interesting to explore in practice is the various possible forms of informative key-variable errors, insofar as the model pertaining to the matched entities in one way or another differs to that of the unmatched entities. Suitable variations of the MEC approach may need to be configured in different situations.

Acknowledgements

The authors thank the associate editor and the reviewers for their constructive comments. Dr. Kim is partially supported by NSF grant MMS 1733572.

Supplementary material

In the supplementary material (arXiv:2009.14797), we present the theoretical convergence property of the proposed algorithm and some special cases of MEC sets for record linkage, and discuss two less practical approaches that can be incorporated into the MEC algorithm. An additional simulation study with low levels of the files' overlap is also presented.

References

Armstrong, J.B., and Mayda, J.E. (1993). Model-based estimation of record linkage error rates. *Survey Methodology*, 19, 2, 137-147. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1993002/article/14459-eng.pdf.

Berger, A.L., Della Pietra, S.A. and Della Pietra, V.J. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics*, 22, 39-71.

Binette, O., and Steorts, R.C. (2020). (almost) all of entity resolution. *arXiv preprint arXiv:2008.04443*.

Christen, P. (2007). A two-step classification approach to unsupervised record linkage. In *Proceedings of the Sixth Australasian Conference on Data Mining and Analytics*, Citeseer, 70, 111-119.

Christen, P. (2008). Automatic record linkage using seeded nearest neighbour and support vector machine classification. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 151-159.

Christen, P. (2012). A survey of indexing techniques for scalable record linkage and deduplication. *IEEE Transactions on Knowledge and Data Engineering*, 24(9), 1537-1555.

Copas, J., and Hilton, F. (1990). Record linkage: Statistical models for matching computer records. *Journal of the Royal Statistical Society, Series A, (Statistics in Society)*, 153(3), 287-312.

Enamorado, T., Fifield, B. and Imai, K. (2019). Using a probabilistic model to assist merging of large-scale administrative records. *American Political Science Review*, 113(2), 353-371.

Fellegi, I.P., and Sunter, A.B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64(328), 1183-1210.

Gull, S.F., and Daniell, G.J. (1984). Maximum entropy method in image processing. *IEE Proceedings 131F*, 646-659.

Hand, D., and Christen, P. (2018). A note on using the f-measure for evaluating record linkage algorithms. *Statistics and Computing*, 28(3), 539-547.

Herzog, T.N., Scheuren, F.J. and Winkler, W.E. (2007). *Data Quality and Record Linkage Techniques*. Springer Science & Business Media.

Jaro, M.A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association*, 84(406), 414-420.

Larsen, M.D., and Rubin, D.B. (2001). Iterative automated record linkage using mixture models. *Journal of the American Statistical Association*, 96(453), 32-41.

McLeod, P., Heasman, D. and Forbes, I. (2011). Simulated data for the on the job training. *Essnet DI*, 70. Available at http://www.crosportal.eu/content/job-training.

Newcombe, H.B., Kennedy, J.M., Axford, S. and James, A.P. (1959). Automatic linkage of vital records. *Science*, 130(3381), 954-959.

Nguyen, X., Wainwright, M.J. and Jordan, M.I. (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11), 5847-5861.

Nigam, K., Lafferty, J. and McCallum, A. (1999). Using maximum entropy for text classification. In *IJCAI-99 Workshop on Machine Learning for Information Filtering*, Stockholom, Sweden, 1, 61-67.

Owen, A., Jones, P. and Ralphs, M. (2015). Large-scale linkage for total populations in official statistics. *Methodological Developments in Data Linkage*, 170-200.

Sadinle, M. (2017). Bayesian estimation of bipartite matchings for record linkage. *Journal of the American Statistical Association*, 112(518), 600-612.

Sarawagi, S., and Bhamidipaty, A. (2002). Interactive deduplication using active learning. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 269-278.

Steorts, R.C. (2015). Entity resolution with empirically motivated priors. *Bayesian Analysis*, 10, 849-875.

Stringham, T. (2021). Fast Bayesian record linkage with record-specific disagreement parameters. *Journal of Business & Economic Statistics*, 0(0), 1-14.

Tancredi, A., and Liseo, B. (2011). A hierarchical Bayesian approach to record linkage and population size problems. *The Annals of Applied Statistics*, 5(2B), 1553-1585.

Winkler, W.E. (1988). Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 667-671.

Winkler, W.E. (1993). Improved decision rules in the Fellegi-Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 274-270.

Winkler, W.E. (1994). Advanced methods for record linkage. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 467-472.

Winkler, W.E., and Thibaudeau, Y. (1991). *An Application of the Fellegi-Sunter Model of Record Linkage to the 1990 US Decennial Census*. Citeseer.

Xu, H., Li, X., Shen, C., Hui, S.L. and Grannis, S. (2019). Incorporating conditional dependence in latent class models for probabilistic record linkage: Does it matter? *Annals of Applied Statistics*, 13(3), 1753-1790.

Zhang, G., and Campbell, P. (2012). Data survey: Developing the statistical longitudinal census dataset and identifying its potential uses. *Australian Economic Review*, 45(1), 125-133.

Survey Methodology, June 2022
Vol. 48, No. 1, pp. 25-48
Statistics Canada, Catalogue No. 12-001-X

The anchoring method: Estimation of interviewer effects in the absence of interpenetrated sample assignment

Michael R. Elliott, Brady T. West, Xinyu Zhang and Stephanie Coffey[1]

Abstract

Methodological studies of the effects that human interviewers have on the quality of survey data have long been limited by a critical assumption: that interviewers in a given survey are assigned random subsets of the larger overall sample (also known as interpenetrated assignment). Absent this type of study design, estimates of interviewer effects on survey measures of interest may reflect differences between interviewers in the characteristics of their assigned sample members, rather than recruitment or measurement effects specifically introduced by the interviewers. Previous attempts to approximate interpenetrated assignment have typically used regression models to condition on factors that might be related to interviewer assignment. We introduce a new approach for overcoming this lack of interpenetrated assignment when estimating interviewer effects. This approach, which we refer to as the "anchoring" method, leverages correlations between observed variables that are unlikely to be affected by interviewers ("anchors") and variables that may be prone to interviewer effects to remove components of within-interviewer correlations that lack of interpenetrated assignment may introduce. We consider both frequentist and Bayesian approaches, where the latter can make use of information about interviewer effect variances in previous waves of a study, if available. We evaluate this new methodology empirically using a simulation study, and then illustrate its application using real survey data from the Behavioral Risk Factor Surveillance System (BRFSS), where interviewer IDs are provided on public-use data files. While our proposed method shares some of the limitations of the traditional approach – namely the need for variables associated with the outcome of interest that are also free of measurement error – it avoids the need for conditional inference and thus has improved inferential qualities when the focus is on marginal estimates, and it shows evidence of further reducing overestimation of larger interviewer effects relative to the traditional approach.

Key Words: Clustering; Intraclass correlation; Design effects; Behavioral Risk Factor Surveillance System.

1. Introduction

Despite the best efforts of survey organizations to standardize the training of both face-to-face and telephone survey interviewers (Fowler and Mangione, 1989), numerous researchers have shown that estimates of key population parameters tend to vary between interviewers (e.g., Groves, 2004; Schnell and Kreuter, 2005; West and Olson, 2010; West and Blom, 2017). This variability may be due to verbal or nonverbal signals sent (likely unintentionally) by different interviewers, or by demographic features of the interviewer that reveal interviewer preferences and expectations (West and Blom, 2017). Even simpler factual items and self-administered items have been found to show variation across interviewers, despite the random assignment of respondents to interviewers (e.g., Kish, 1962; Groves and Magilavy, 1986; O'Muircheartaigh and Campanelli, 1998).

This intra-interviewer correlation, generally referred to as an *interviewer effect*, reduces the efficiency of survey estimates and decreases effective sample sizes given fixed survey costs in a manner similar to

1. Michael R. Elliott, University of Michigan Institute for Social Research, 426 Thompson St., Ann Arbor, MI 41809, University of Michigan Department of Biostatistics, School of Public Health, 1415 Washington Heights, Ann Arbor, MI 48109. E-mail: mrelliot@umich.edu; Brady T. West and Xinyu Zhang, University of Michigan Institute for Social Research, 426 Thompson St., Ann Arbor, MI 41809; Stephanie Coffey, US Census Bureau, 4600 Silver Hill Rd, Suitland-Silver Hill, MD 20746.

cluster sampling, due to the presence of a common effect across subjects that induces correlation. It can be conceptualized in statistical terms as a random effect common to all observations obtained by a given interviewer, whose variance is termed "interviewer variance". Accounting for this variance is critical to get correct statistical inference. In addition, as part of data collection monitoring, survey managers can use unbiased estimates of interviewer effects to identify interviewers that are having extreme effects on particular survey outcomes in real time and may need additional training to curb inappropriate behaviors.

A key assumption in the estimation of interviewer variance – whether via random effects models, or indirectly through use of generalized estimation equation/Taylor Series approaches – is interpenetrated sampling, or the random assignment of sampled cases to interviewers. Thus Schnell and Kreuter (2005) estimate interviewer effects in a face-to-face survey where interviewers are nested within PSUs and respondents within a PSU are randomly assigned to an interviewer, while O'Muircheartaigh and Campanelli (1998) use a cross-classified model in a design where respondents are randomly assigned to interviewers who worked in multiple PSUs. Interpenetrated sampling helps to ensure unbiased estimation of interviewer variance by ensuring there is no "spurious" variance introduced by certain types of respondents being more likely to be assigned to a given interviewer (e.g., older respondents being associated with interviewers working during the day), just as randomization ensures unbiased estimation of treatment effects in clinical trials. Unfortunately, interpenetrated sampling is logistically infeasible in many sample designs.

Recent studies of interviewer variance have adopted ad-hoc analytic approaches to "adjusting" for the effects of selected covariates that may introduce spurious correlation within interviewers based on non-interpenetrated sample designs (e.g., covariates describing features of sampling areas), claiming that any remaining variance in survey estimates across interviewers is mostly attributable to the interviewers (West and Blom, 2017). While this approach may in principle work to reduce spurious correlations between interviewers and outcomes if such covariates are available, it comes at the price of requiring conditional inference for the substantive variable of interest. This is particularly problematic if our goal is inference that properly accounts for interviewer effects in variance estimation without inappropriately adjusting for covariates that are not of interest. For example, if our interest is in the mean of a survey variable Y, $E(Y)=\mu$, while appropriately accounting for the additional variance introduced by "clustering" from multiple interviewers conducted by a single interviewer, adjusting for multiple covariates $(X_1,\ldots,X_p)$ yields an estimator of β_0 under the model $E(Y)=\beta_0+\sum_{k=1}^{p}\beta_k x_k$. It is clear that $\mu\neq\beta_0$ unless either $\beta_1=\ldots=\beta_p=0$ (in which case there cannot be adjustment for spurious correlations between interviewers and outcome), $E(X_1)=\ldots=E(X_p)=0$, or there is some extremely unlikely cancellation of regression components. (For readers familiar with causal inference, this is somewhat analogous to marginal structural models (MSMs), which avoid using confounders in a regression model while still accounting for confounding, Joffe, Ten Have, Feldman and Kimmel (2004), although our approach is fully model-based rather than model-assisted as in MSMs.) While centering the covariates can guarantee the second condition in the absence of interactions, this is not always desirable or noted, and even if doable may not

leave the remaining residuals with the desired distributional characteristics. With the present study, we aim to provide survey researchers with a means to estimate interviewer variance (either to improve the quality of estimates or inform survey operations) in the absence of interpenetration without conditioning on covariates in the traditional manner.

Our approach, which we refer to as the "anchoring" method, leverages correlations between observed variables that are unlikely to be affected by interviewers ("anchors") and variables that may be prone to interviewer effects (e.g., sensitive or complex factual questions) to statistically remove components of within-interviewer correlations that a lack of interpenetrated assignment may introduce. The improved estimates of interviewer effects on survey measures will increase the ability of survey analysts to correct estimates of interest for interviewer effects, and enable survey managers to adaptively manage a data collection in real time and intervene when particular interviewers are producing survey outcomes that vary substantially from expectations.

In Section 2, we provide some background on the important problem of interviewer variance, as well as a discussion of its estimation and impact on inference. In Section 3, we introduce the anchoring method and its development in a frequentist and Bayesian framework, as well as the heuristic interpretation and issues related to choice of variables. In Section 4 we empirically evaluate the properties of this new method using a simulation study, and in Section 5 we illustrate the method using real data from the Behavioral Risk Factor Surveillance System (BRFSS). In Section 6 we provide concluding remarks as well as some discussion of implementation and monitoring of the method in practise.

2. Background

2.1 Interviewer variance

Between-interviewer variance affects survey estimates in a manner similar to the design effects introduced by cluster sampling. One can estimate the multiplicative increase in the total variance of an estimated mean as $\text{deff} = 1 + \rho_{\text{int}}(m-1)$, where m is the average number of interviews conducted by individual interviewers and ρ_{int} is the within-interviewer correlation in answers elicited to a particular survey question (Kish, 1965). Typical values of 35 respondents per interviewer and 0.03 for ρ_{int} would therefore *double* the estimated variance of the mean, relative to the variance with ρ_{int} equal to zero. Failure to account for the within-interviewer correlation introduced by interviewer effects leads to *misspecification effects* (Skinner, Holt and Smith, 1989), resulting in anti-conservative inference due to underestimation of standard errors.

2.2 Estimation of interviewer variance

Researchers may wish to estimate interviewer variance for correct statistical inference (Elliott and West, 2015), to identify interviewers having unusual effects on data collection outcomes for purposes of responsive survey design, or as the focus of a methodological study designed to reduce its impact by

understanding its causes (e.g., Brunton-Smith, Sturgis and Williams, 2012; Sakshaug, Tutz and Kreuter, 2013). Interpenetrated designs, which assign sampled cases to interviewers at random, allow for interviewer variance to be accounted for using standard methods that account for clustering in the observed data: generalized estimating equations (Liang and Zeger, 1986) or mixed-effects models (Laird and Ware, 1982; Stiratelli, Laird and Ware, 1984). Temporarily ignoring sampling weights, a simple model for a normally-distributed variable of interest that accounts for interviewer variance is

$$Y_{ijk} = \mu + a_i + b_{ij} + \varepsilon_{ijk},\ a_i \sim N\left(0, \sigma_a^2\right),\ b_{ij} \sim N\left(0, \sigma_b^2\right),\ \varepsilon_{ijk} \sim N\left(0, \sigma^2\right), \tag{2.1}$$

where i indexes a primary sampling unit (PSU), j indexes the interviewer within the ith PSU, and k the respondent associated with the jth interviewer in the ith PSU. Assuming that all of the error terms are independent, that there are an average of J interviewers in each of the I PSUs, and that there are an average of K interviews per interviewer, the variance of the mean estimator $\hat{\mu} = \bar{y}$ is approximately inflated by a factor of $1 + \rho_a\left(JK - 1\right) + \rho_b\left(K - 1\right)$, where $\rho_a = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_b^2 + \sigma^2}$ and $\rho_b = \frac{\sigma_b^2}{\sigma_a^2 + \sigma_b^2 + \sigma^2}$. As a practical matter, when the variance of $\hat{\mu}$ is the only quantity of interest, the second stage of clustering due to an interviewer can be ignored, as in an "ultimate cluster" design (Kalton, 1983). Treating the random effect of the PSU as $\tilde{a}_i = a_i + \sum_{j=1}^{J} b_{ij}$ with variance $\sigma_{\tilde{a}}^2 = \sigma_a^2 + J\sigma_b^2$, the variance of the mean estimator $\hat{\mu}$ is inflated by a factor of $1 + \rho_{\tilde{a}}\left(JK - 1\right)$, where $\rho_{\tilde{a}} = \frac{\sigma_{\tilde{a}}^2}{\sigma_{\tilde{a}}^2 + \sigma^2}$.

If multiple interviewers are nested within a single PSU as assumed in (2.1), interviewer variances can still be estimated for methodological purposes using multistage hierarchical linear models. However, for reasons of cost efficiency, many area probability samples require a given interviewer to restrict their efforts to a single sampling area (e.g., the U.S. National Survey of Family Growth; see Lepkowski, Mosher, Groves, West, Wagner and Gu, 2013), which completely aliases the components of variance due to interviewers and areas. Such designs preclude any type of direct estimation of interviewer variance, although from a purely analytic perspective, accounting for clustering using the PSU IDs in analysis will account for the additional interviewer variance introduced.

For other types of surveys – and in particular telephone surveys – this "automatic" accommodation of interviewer effects at the variance estimation stage afforded by "ultimate cluster" approaches does not occur. A spectacular example of this is the Behavioral Risk Factor Surveillance System (BRFSS; Centers for Disease Control, 2013), a massive annual telephone survey sponsored by the Centers for Disease Control that is the only Federal health survey designed to provide state-level estimates of key health factors such as smoking rates, obesity measures, and cancer screening. Elliott and West (2015) found no evidence that any substantial proportion of the 1,000 + manuscripts published using BRFSS data accounted for interviewer effects when conducting variance estimation based on these data, despite variance inflation factors of 10 or more at the state level for estimates such as mean self-rated health. These authors found evidence of substantial interviewer effects for selected survey items, and variability in the variance of these effects themselves across states, when applying both model-based and design-based approaches to estimate the variance (although this analysis used naïve estimators in contrast to

either the standard regression or the anchoring methods discussed here, and so may have overestimated this variance).

Importantly, secondary analysts still do not know for sure if these components of variance are arising due to sampling variability, true measurement error introduced by the interviewers, or differential non-response among the interviewers. Because of the design effect definition noted above, their impact on inference can still be large even if the intra-class correlation (ICC) is small or moderate, since interviewers typically conduct many interviews. Thus when Groves and Magilavy (1986) found mean ICCs between 0.002 and 0.02 among 25 to 55 variables across each of nine telephone surveys of political, health, and economic issues, the design effect would range between 1.04 and 1.38 for studies in which interviewers average 20 interviews each, and between 1.10 and 1.98 if interviewers average 50 interviews each. Some outcomes can have much higher ICCs – Cernat and Sakshaug (2021) found ICCs on the order of 0.30 for biometric measures, which would yield design effects on the order of 15 if 50 interviews were conducted per interviewer. Although interviewer variance studies for face-to-face data collections tend to be rare because interpenetrated sample designs are more difficult to implement in such settings, Schnell and Kreuter (2005) found a median overall design effect of 2.0 in a multi-stage sample survey on fear of crime, which was mostly attributable to interviewer effects rather than spatial clustering. Thus the need for analysts to accommodate interviewer effects is clear.

2.3 Accounting for interviewer variance in inference in the absence of interpenetration

As noted in Section 2.2, when interviewers are nested within PSUs, standard methods of variance estimation based on "ultimate clusters" (Kalton, 1983) that account for the dependence of observations within a PSU will "automatically" absorb measurement error due to interviewers into the within-PSU correlation. However, whenever interviewers are not nested within PSUs – as can occur in some area probability samples where interviewers cross sampling unit segments (e.g., O'Muircheartaigh and Campanelli, 1998; Vassallo, Durrant and Smith, 2017) – clustering induced by interviewer effects must be accounted for directly. In such situations, cross-classified random effects models (Rasbash and Goldstein, 1994) of the form

$$E\left(Y_{hij}\right)=\theta+a_h+b_i,\ a_h\sim N\left(0,\tau_a^2\right),\ b_i\sim N\left(0,\tau_b^2\right) \tag{2.2}$$

may be employed, where h indexes PSUs, i indexes interviewers, and j indexes interviews conducted by the ith interviewer (e.g., O'Muircheartaigh and Campanelli, 1998; Schnell and Kreuter, 2005; Biemer, 2010; Durrant, Groves, Staetsky and Steele, 2010). Extensions of these models are also possible for non-linear link functions using generalized linear mixed models (e.g., Vassallo et al., 2017).

Unfortunately, interpenetration can fail, either due to differential non-response error among interviewers (West and Olson, 2010; West, Kreuter and Jaenichen, 2013), non-random shift assignment (e.g., with daytime interviewers more likely to interview non-working respondents), or other common

practices used to increase response rates, such as assigning experienced interviewers to more difficult respondents (Brunton-Smith et al., 2012). In the absence of interpenetration, standard methods to account for interviewer variance may lead to "spurious" correlations within interviewers that have nothing to do with interviewer-induced measurement error.

The literature is not completely devoid of approaches for estimating (and accommodating) interviewer variance in non-interpenetrated sample designs. Fellegi (1974), Biemer and Stokes (1985), Kleffe, Prasad and Rao (1991), and Gao and Smith (1998) developed statistical methods for area probability samples that assumed interpenetration for a random subset of PSUs, and a single interviewer in each of the remaining PSUs. More recent work has considered methods for estimation of interviewer variance in binary survey variables in related settings of *partial interpenetration* (von Sanden and Steel, 2008). Rohm, Carstensen, Fischer and Gnambs (2021) used a two-parameter item response theory model to separate area and interviewer effects under this assumption, which de-confounds interviewer and area effects to the extent that each interviewer recruits in multiple areas and vice versa (although lack of random assignment within an area can still yield some degree of variance component bias). These methods are useful for obtaining estimates of interviewer variance separate from area homogeneity for purposes of assessing the independent impact of such variance. However, they are not relevant for our more general setting of interest, where interviewers may not cross PSUs and are not working random subsamples of the full sample (i.e., no interpenetration).

Another common method found in the literature for grappling with the problem of non-interpenetrated sample designs when estimating interviewer variance is adjustment for the effects of respondent- and area- or interviewer-level covariates in multilevel models (Hox, 1994; Schaeffer, Dykema and Maynard, 2010; West, Kreuter and Jaenichen, 2013). These methods are largely ad-hoc, and rely on the assumption that the included covariates adequately account for all sources of variability that arise from the areas (and would thus be attributed to the interviewers if the covariates were not accounted for). This approach suffers from two major shortcomings. First, many studies, and especially those relying on publicly available data, may not contain sufficient area- or interviewer-level covariate information to adequately account for the lack of randomization in interviewer assignment. Second, the resulting estimators condition on these covariates, and these conditional estimators are typically not of interest, with the focus being on either marginal estimates of descriptive parameters, such as means or totals, or parameters in models that typically do not condition on (or include) covariates.

3. The anchoring method

As noted in Section 2.3, existing methods adjust for possible interviewer effects introduced at the recruitment and measurement stages of data collection by including respondent- and area- or interviewer-level covariates in multilevel models (Hox, 1994), but such adjustment may be erroneous if part of the interviewer variance is simply arising due to non-interpenetrated sampling. As noted by Elliott and West

(2015), if subjects with similar values on a variable of interest are assigned to interviewers in a non-random fashion – for example, if a telephone interviewer working day shifts tends to interview older respondents, where age may be correlated with main variables of interest – these variables will be correlated with specific interviewers. However, we are just re-ordering the random sample, not introducing measurement error in the manner described in Section 1, e.g., West and Blom (2017). Thus the actual data are not being altered, and there are no true interviewer effects: we term the resulting within-interviewer correlation "spurious" from a variance inflation perspective. Thus estimating interviewer effects while failing to account for differential sample assignment can lead to conservative inferences, resulting in misleadingly large estimates of interviewer variance, p-values and confidence intervals that are too wide, and incorrect operational decisions based on predicted effects for individual interviewers.

To address this important gap in the literature, we describe an "anchoring" method that analysts can use to estimate the unique components of variance due to interviewer effects on selection and measurement. The method aims to leverage correlations between variables where interviewer measurement error is of concern and variables known – or reasonably believed – to be free of measurement error to remove the fraction of the within-interviewer correlation that is due to non-interpenetrated sample assignment. In the simplest case, if we have two variables, one (Y_1) treated as measurement error-free (the "anchor") and one (Y_2) treated as possibly having interviewer-induced measurement error, and our objective is to estimate a mean of Y_2, we fit a multilevel model to the observed data for the two variables that includes a random interviewer effect only for the variable subject to measurement error:

$$y_{ijk} = \mu_k + I(k=2)\, b_i + \varepsilon_{ijk}. \tag{3.1}$$

In (3.1), $i = 1, \ldots, I$ indexes interviewers, $j = 1, \ldots, J_i$ indexes respondents within interviewers, $k = 1, 2$ indexes the variable (1 = anchor, 2 = variable of interest), $b_i \sim N\left(0, \sigma_b^2\right)$ is the interviewer effect, and

$$\begin{pmatrix} \varepsilon_{ij1} \\ \varepsilon_{ij2} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \right).$$

Our focus of inference in this manuscript is μ_2, although σ_b^2 or b_i may also be of interest if the focus is on interviewer variance or determining individual interviewers who are contributing to that variance.

To provide a heuristic explanation of why this proposed "anchoring" approach works, assume that y_{ij1} and y_{ij2} net of b_i are almost perfectly correlated. Since y_{ij1} lacks measurement error, it can serve as a proxy for the non-measurement error component of y_{ij2}, absorbing artificial error in y_{ij2} induced in the ordering of the data. Lack of interpenetration means that estimating a linear mixed model using y_{ij2} only will yield an upwardly biased estimate of σ_b^2. If $\sigma_{12} > 0$, information will be available to reduce the bias in $\hat{\sigma}_b^2$, with large samples and high correlations between ε_{ij1} and ε_{ij2} yielding increasingly accurate estimates of σ_b^2 and thus of the true impact of the interviewer-induced measurement error on the variance of $\hat{\mu}_2$.

This approach easily extends to the setting where $K-1 \geq 2$ "anchoring" variables free from measurement error are available:

$$y_{ijk} = \mu_k + I\left(k = K\right) b_{iK0} + \varepsilon_{ijk},\ i = 1, \ldots, I,\ j = 1, \ldots, J_i, k = 1, \ldots, K. \tag{3.2}$$

In this case, the first $K-1$ variables are assumed to be free of interviewer measurement error and the Kth variable is the variable of interest, $b_{iK0} \sim N\left(0, \tau^2\right)$, and $\left(\varepsilon_{ij1} \ldots \varepsilon_{ijK}\right)^T \sim N_K\left(0, \Sigma\right)$, where Σ is an unstructured $K \times K$ covariance matrix. Alternatively, instead of using (3.2) directly, we can reduce (3.2) back to the bivariate setting in (3.1) by replacing Y_{1i} with the best linear predictor of Y_{Ki} using the anchoring variables: $\hat{Y}_{Ki} = E\left(Y_{Ki} \middle| Y_{1i}, \ldots, Y_{K-1,i}\right) = \hat{\boldsymbol{\beta}}^T \mathbf{X}_i$ where $\hat{\mathbf{X}}_i = \left(Y_{1i}, \ldots, Y_{K-1,i}\right)^T$ and $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{Y}_K$.

3.1 Estimation remarks

One can use standard linear mixed model software (e.g., SAS PROC MIXED) to fit the models in (3.1) or (3.2) and obtain a restricted maximum likelihood (REML) point estimate $\hat{\mu}_2$ together with an associated variance estimate. We have provided an annotated example of such code in the supplemental materials. Weights used to account for unequal probabilities of selection, non-response adjustment, and calibration to known population values can be incorporated using pseudo-maximum likelihood estimation (PML; Pfeffermann, Skinner, Holmes, Goldstein and Rasbash, 1998; Rabe-Hesketh and Skrondal, 2006) when fitting the models in (3.1) or (3.2). We would generally recommend that interviewers be assigned a weight of 1 when fitting weighted multilevel models of these forms, to mimic the notion of simple random sampling of interviewers from a hypothetical population of interviewers. The weights for respondents should be rescaled to sum to the final respondent count for each interviewer (Carle, 2009), and extensions of the PML method outlined by Veiga, Smith and Brown (2014) and Heeringa, West and Berglund (2017, Chapter 11) can be used to incorporate the rescaled weights in estimation of the residual covariance structure in (3.1) or (3.2). In multistage samples where interviewers cross geographic areas, cross-classified random effects models (Rasbash and Goldstein, 1994) can also be utilized.

3.2 The Bayesian anchoring method

In the presence of prior information on the parameters of interest in this model (e.g., in a repeated cross-sectional survey using interviewer administration), the models in (3.1) or (3.2) can also be fitted using a Bayesian approach to incorporate the prior information. In repeated surveys that carefully monitor interviewer performance, good predictions of individual interviewer effects based on the estimated variance component are important. Given historical data from a survey with the same essential design conditions, one can estimate the parameters of interest in (3.1) using this historical data, and then define informative prior distributions for these parameters. (Examples of these types of surveys would include high-quality government sponsored surveys with repeated cross-sectional data collections, such as the National Health Interview Survey or, for the example considered in this paper, the Behavioral Risk Factor

Surveillance System.) Specifically, we consider a prior distribution on the interview effect standard deviation σ_b that follows a half t distribution (Gelman, 2006) with degrees of freedom ν and standard deviation s:

$$p\left(\sigma_b \mid \nu, s\right) = \frac{2\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi s^2}}\left(1+\frac{\sigma_b^2}{\nu s^2}\right)^{-\frac{\nu+1}{2}}, \quad \tau \geq 0. \tag{3.3}$$

Following Gelman, we assume $\nu = 3$, and we estimate s based on prior estimates of interviewer effects. We consider standard weak priors for the fixed effect means: $p\left(\mu_k\right) \overset{\text{ind}}{\sim} N\left(0, 10^6\right)$ and for the residual variance:

$$p\begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \sim \text{INV} - \text{WISHART}\left(2, I\right).$$

This approach offers advantages relative to likelihood ratio testing approaches that rely on asymptotic theory, particularly for smaller samples. By using prior information to constrain the resulting posterior distribution for the interviewer variance components, it generally prevents extremely large draws of the variance component while not constraining the means or residual variances. It also constrains posterior draws of variance components to be greater than zero, enabling inference based on small components of variance, while frequentist model-fitting procedures generally fix such estimates of variance components to be exactly zero (which equates to a rather unreasonable assumption that each interviewer produces exactly the same survey estimate; West and Elliott, 2014). In such cases, the effects of interviewers (even if they are small) would be ignored completely; the Bayesian approach would still enable small effects to be integrated into the inference. The Bayesian approach also yields credible intervals for the interviewer variance components based on posterior draws.

3.3 Choosing anchoring variables

A key assumption underlying both the standard regression-based approach and the "anchoring" method is that selected variables are free from interviewer-induced error. Like the "missing at random" assumption in the missing data literature, we do not expect that there will often be cases where we can be certain of this, but that approximations may be available based on simple demographic measures (e.g., age) or other factual questions with simple response options (e.g., current employment) and little room for the introduction of interviewer error. The identification of error-free covariates in advance of data collection is an important substantive and methodological component of this approach, and prior methodological literature on the variables most prone to interviewer effects (West and Blom, 2017) can be consulted for this component of the approach.

As we note above, if we have multiple error-free covariates measured on the respondents, we can preserve their predictive power (and thus the correlation of the anchor's residuals with the residuals of the variable of interest) by computing a linear predictor of the variable of interest from a linear model that includes fixed effects of all of the error-free covariates. We consider such an approach in our simulation

studies and applications, and compare it with the "standard" approach of simply adjusting for these covariates in a multilevel model in an effort to improve the estimate of the interviewer variance component (Hox, 1994).

Finally, the anchoring approach employs mixed-effects models that should yield correct estimates with a sufficient amount of data. However, these models may be more difficult to fit, especially for smaller samples, and we therefore also consider alternative Bayesian approaches when evaluating the anchoring approach.

4. Simulation study

We first consider an empirical simulation study of the proposed "anchoring" approach. We repeatedly simulated samples of data from a quadrivariate normal distribution, $\left(Y_{1ij}^{*}\ Y_{2ij}^{*}\ Y_{3ij}^{*}\ Z_{ij}\right) \sim N_4\left(\boldsymbol{\mu}, \Sigma\right)$, where $j = 1, \ldots, J = 30$ indexes hypothetical respondents nested within $i = 1, \ldots, I = 30$ interviewers, $Y_{kij(z)} = Y_{kij(z)}^{*} + I\left(k = 3\right) b_i$ for $b_i \sim N\left(0, \sigma_b^2\right)$, and $Y_{kij(z)}^{*}$ is ordered by the values of Z_{ij} prior to assignment of respondents to the 30 interviewers. $Y_{1ij(z)}$ and $Y_{2ij(z)}$ are the observed values without interviewer-induced measurement error, while $Y_{3ij(z)}$ is observed with interviewer-induced measurement error, and Z_{ij} is a (nuisance and unobserved) covariate that induces extraneous variability when the design is treated as interpenetrated. (One might think of Y_1 and Y_2 as measurement-error free demographic variables and Y_3 as a continuous self-reported overall health measure, which is potentially prone to interviewer effects, and Z as amount of time spent at home, which is associated with interviewer scheduling by shift.)

Given this data generating model, we note that a higher correlation of Z with the other measurements will introduce what appears to be interviewer variance because of the ordering of $Y_{kij(z)}^{*}$ by the values of Z_{ij} above and beyond the true random interviewer effects on Y_2 (given by b_i). This is the lack of interpenetrated assignment that we wish to adjust for with the proposed anchoring method, which aims to isolate the unique interviewer variance σ_b^2 that does not arise from simple assignment of cases to interviewers. For simplicity, we assume that $\mu_{Y_1} = \mu_{Y_2} = \mu_{Y_3} = \mu_z = \mu,\ \sigma_{Y_1}^2 = \sigma_{Y_2}^2 = \sigma_{Y_3}^2 = \sigma_Z^2 = 1$ and $\rho_{Y_1Y_2} = \rho_{Y_1Y_3} = \rho_{Y_1Z} = \rho_{Y_2Y_3} = \rho_{Y_2Z} = \rho_{Y_3Z} = \rho.$

We consider four models used to estimate the mean of Y_3 and the associated interviewer effect variance:

Unadjusted: $Y_{3ij} \sim N\left(\mu_3 + b_i, \sigma_3^2\right)$

Adjusted: $Y_{3ij} \sim N\left(\mu_3 + \beta_1 y_{1ij} + \beta_2 y_{2ij} + b_i, \sigma_3^2\right)$

Anchoring: $\begin{pmatrix} Y_{1ij} \\ Y_{2ij} \\ Y_{3ij} \end{pmatrix} \sim N_3\left(\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 + b_i \end{bmatrix}, \Sigma\right)$

Anchoring-Linear Predictor: $\begin{pmatrix} \hat{Y}_{3ij} \\ Y_{3ij} \end{pmatrix} \sim N_2 \left(\begin{bmatrix} \mu_1 \\ \mu_2 + b_i \end{bmatrix}, \Sigma \right)$

where y_{kij} is the observed realization of Y_{kij}, $b_i \sim N\left(0, \sigma_b^2\right)$, and, in the anchoring-linear predictor model, $\hat{Y}_{3ij} = \hat{\beta}_0 + \hat{\beta}_1 y_{1ij} + \hat{\beta}_2 y_{2ij}$ where $\hat{\boldsymbol{\beta}}$ is obtained from the linear regression of Y_3 on Y_1 and Y_2. We estimate the mean of Y_3 as the REML estimator of μ_3 and similarly the associated interviewer effect variance as the REML estimator of σ_b^2.

We consider the power to reject the null hypothesis that the mean of the observed variables is zero (at the 0.05 level) and the empirical bias in the estimation of the variance of the random interviewer effects, σ_b^2. We evaluated the empirical bias by computing the difference between the mean of the simulated estimates of the variance component and the true value of the variance component specific to a given simulation scenario. Our simulation study design considers a full factorial design where $\mu = \{0, 0.5\}$, $\rho = \{0.25, 0.5, 0.75\}$, and $\sigma_b^2 = \{0.1, 0.5, 0.9\}$. We generated 200 independent simulations for each of the 18 cross-classifications of values on these parameters. Table 4.1 presents the results of the simulation study.

Table 4.1
Results of the empirical simulation study. Best performing method italicized (note that when $\mu = 0$, ideal power if 0.05)

True values of model parameters			Power: $H_0 : \mu = 0$ *vs.* $H_A : \mu \neq 0$				Empirical Bias of $\hat{\sigma}_b^2$			
μ	ρ	σ_b^2	Unadjusted	Adjusted	Anchor	Anchor-Linear Predictor	Unadjusted	Adjusted	Anchor	Anchor-Linear Predictor
0	0.25	0.1	0.03	*0.04*	*0.04*	*0.04*	0.063	0.029	*0.027*	*0.027*
0	0.25	0.5	0.03	0.08	*0.04*	*0.04*	0.070	*0.022*	0.033	0.032
0	0.25	0.9	0.07	*0.04*	*0.06*	*0.06*	0.078	*0.037*	0.044	0.043
0	0.5	0.1	0.00	*0.03*	0.02	0.02	0.255	0.061	*0.056*	*0.056*
0	0.5	0.5	0.01	*0.04*	0.03	0.03	0.247	0.058	*0.054*	*0.053*
0	0.5	0.9	0.02	*0.04*	*0.04*	*0.04*	0.251	*0.049*	0.061	0.061
0	0.75	0.1	0.00	*0.02*	0.01	0.01	0.568	*0.074*	0.078	0.076
0	0.75	0.5	0.00	0.04	*0.05*	*0.05*	0.555	0.098	*0.084*	*0.084*
0	0.75	0.9	*0.04*	*0.04*	*0.06*	*0.06*	0.602	*0.099*	0.103	0.103
0.5	0.25	0.1	*1.00*	*1.00*	*1.00*	*1.00*	0.069	*0.025*	0.032	0.032
0.5	0.25	0.5	*0.96*	0.68	*0.96*	*0.96*	0.072	0.044	*0.034*	*0.034*
0.5	0.25	0.9	*0.76*	0.48	0.75	0.75	0.075	0.040	*0.039*	*0.037*
0.5	0.5	0.1	*1.00*	0.87	*1.00*	*1.00*	0.261	0.062	0.062	*0.061*
0.5	0.5	0.5	0.92	0.44	*0.96*	*0.96*	0.269	*0.062*	0.067	0.067
0.5	0.5	0.9	0.75	0.24	*0.80*	*0.80*	0.248	0.068	0.064	*0.063*
0.5	0.75	0.1	*1.00*	0.62	*1.00*	*1.00*	0.567	0.079	0.078	*0.077*
0.5	0.75	0.5	0.81	0.27	*0.96*	*0.96*	0.507	0.103	*0.082*	*0.082*
0.5	0.75	0.9	0.58	0.22	*0.70*	*0.70*	0.598	*0.100*	0.106	0.106

Several notable patterns emerge from the simulation results in Table 4.1. First, as the values of ρ increase, the anchoring method produces larger reductions in the overestimation of interviewer variance relative to the unadjusted model. Recall that this was expected by design, given the initial ordering of the observations by Z prior to assignment to interviewers, which introduces artificial variance among the interviewers. Similarly, as anticipated, estimation of the interviewer variance using covariate adjustment is similar to the anchoring method when this variance is not large, although there is evidence of a somewhat larger reduction in bias when the variance is large.

In addition, for the non-zero values of μ, higher values of ρ yield larger improvements in power when using the anchoring method when compared with the unadjusted estimator, since more of the extraneous variance is correctly allocated. Both the unadjusted and an anchoring method yield higher power than the adjusted estimator, since the adjusted estimator is biased for non-zero means of Y_{1ij} and Y_{2ij} when they are correlated with Y_{3ij}. Smaller values of ρ approximate an interpenetrated design, and as a result, the unadjusted estimation approach does not produce substantially different results from the adjusted or anchoring approach. The empirical bias in the estimation of σ_b^2 is unrelated to the value of σ_b^2 but is entirely a function of ρ, since that drives the spurious within-interviewer correlation due to the unobserved Z. Finally, we note that replacing the actual values of Y_{1ij} and Y_{2ij} with a summary measure based on their linear prediction of Y_{3ij} yields virtually identical results to their direct use in the anchoring method. This is partly a function of their common normality; we discuss this limitation in the Discussion section below.

5. Application to the Behavioral risk factor surveillance system

To further illustrate the implementation of our proposed approach, we analyze data from the 2011 and 2012 Behavioral Risk Factor Surveillance System (BRFSS; https://www.cdc.gov/brfss/index.html). The BRFSS is a major national health survey in the U.S. that employs interviewer administration via the telephone, and is one of the few national surveys that provides data users with interviewer identification variables in the public-use versions of its data sets (Elliott and West, 2015). This enables the estimation of interviewer variance components for any BRFSS measures. We only use data from the publicly available data files for these two years in this study.

For illustration purposes, we consider the case where the variable of interest (Y_2) is perceived health status (1 = poor, …, 5 = excellent). We define an "anchoring" variable (Y_1) as the linear predictor of perceived health status from a linear regression model fitted using OLS that includes age, an indicator of obtaining a college degree, an indicator of being a female, and an indicator of white race/ethnicity as covariates. We chose these respondent-level covariates for this application for three reasons: 1) we believe that they are likely to be reported with minimal differential measurement error among interviewers (West and Blom, 2017); 2) they are associated with interviewer assignment, as telephone interviewers tend to work calling shifts at different times of the day, and interview time of day is associated with age and

education (e.g., older respondents and respondents with lower levels of education may be more likely to be interviewed during the day); and 3) they also tend to be correlated with perceived health status (Franks, Gold and Fiscella, 2003).

As part of the application, we also compare the ability of the anchoring method based on this linear predictor to reduce estimates of variance components to that of the more "standard" method that is often used in practice: simply adjusting for these respondent-level covariates in a multilevel model, in an effort to adjust for the fixed effects of these covariates when evaluating the interviewer variance component (Hox, 1994). We make two remarks about this approach, specifically with respect to this application:

1. Centering of the covariates at their means (whether they are binary or continuous) is critical to this approach if inference is focused on the mean of Y_2, as a failure to do this will lead to biased "conditional" estimates of the mean on that variable that depend on the values of the covariates (rather than the overall mean). This is not relevant for the anchoring method.
2. In some cases interviewer-level covariates could be expected to explain more of the artificial interviewer variance due to non-interpenetrated assignment than respondent-level covariates (e.g., area-level socio-demographic information; Hox, 1994; West and Blom, 2017). However, the BRFSS does not provide any interviewer-level covariates.

5.1 Frequentist approach

We considered both frequentist and Bayesian approaches in our analysis, and performed separate analyses of the BRFSS data from each of the 50 states and the District of Columbia for each approach. We only retained cases with complete data on all analysis variables of interest to ensure a common case base no matter the type of analysis being performed. First, in the frequentist approach, we started by estimating means of self-reported health from a given state that assumed independent and identically distributed (i.i.d.) data (i.e., ignoring random interviewer effects):

$$Y_{ij2} = \mu_2 + \varepsilon_{ij2}, \quad \varepsilon_{ij2} \sim N\left(0, \sigma_2^2\right). \tag{5.1}$$

We then fit a "naïve" mixed-effects model including random interviewer effects (of the form in (2.1) but without random PSU effects, given the absence of PSUs in the BRFSS design) to the self-reported health data (ignoring the other covariates), assuming interpenetrated sample assignment within each state:

$$Y_{ij2} = \mu_2 + b_i + \varepsilon_{ij2}, \quad b_i \sim N\left(0, \sigma_b^2\right). \tag{5.2}$$

We estimated the interviewer variance component based on this model and tested the variance component for significance using a mixture-based likelihood ratio test (West and Olson, 2010). We also evaluated the ratio of the estimated variance of mean self-reported health when naively accounting for the interviewer effects to the variance of the mean assuming simple random sampling (i.e., i.i.d. data). The

literature generally refers to this ratio, shown in (5.3), as an "interviewer effect" on a particular descriptive estimate:

$$\text{IntEff}_{\text{naive}} = \frac{\text{var}_{\text{naive}}\left(\hat{\mu}_2\right)}{\text{var}_{\text{iid}}\left(\hat{\mu}_2\right)}. \tag{5.3}$$

Next, after fitting a linear regression model to the perceived health status variable and computing the linear predictor of perceived health status based on the estimated coefficients (denoted in (5.4) by y_{ij1}), we fit the model in (3.1) to implement the anchoring approach:

$$\begin{aligned} y_{ij1} &= \mu_1 + \varepsilon_{ij1} \\ y_{ij2} &= \mu_2 + b_i + \varepsilon_{ij2} \\ b_i &\sim N\left(0, \sigma_b^2\right) \\ \begin{pmatrix} \varepsilon_{ij1} \\ \varepsilon_{ij2} \end{pmatrix} &\sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}\right). \end{aligned} \tag{5.4}$$

Here $Y_{ij1} = \hat{\beta}_0 + \sum_p \hat{\beta}_p x_{ip}$, where $\hat{\boldsymbol{\beta}}$ is obtained from the linear regression of the p anchoring covariates (of which there are four in this application). We then computed the same ratio in (5.3) based on the anchoring approach, where anchoring would be expected to reduce the bias in the estimate of the interviewer effect that would be arising from the naïve approach.

Next, we fitted a model representing the "standard" adjustment approach (Hox, 1994) as follows:

$$Y_{ij2} = \mu_2 + \sum_p \beta_p x_p + b_i + \varepsilon_{ij2},\ b_i \sim N\left(0, \sigma_b^2\right). \tag{5.5}$$

In (5.5), the x_p represent the centered respondent-level covariates indexed by p (the same four anchoring covariates as in (5.4)), with corresponding fixed effects. We once again computed the ratio in (5.3) representing the estimated interviewer effect for comparison with the other approaches. To keep the focus on the potential reduction in bias in the estimation of the interviewer effect, we ignored sampling weights in these analyses.

5.2 Bayesian approach

Next, in the Bayesian approach, we applied the same types of comparative analyses to evaluate the anchoring method, varying whether prior information about the interviewer variance component from the 2011 BRFSS was used (yes or no). This prior information came from implementing the anchoring approach with the same linear predictor in 2011 to determine a prior estimate of the interviewer variance component. In all cases, we assumed non-informative prior distributions for the fixed effects (which recall from (3.1) define the means of the two variables) and the residual variances and covariances in the models.

We defined an informative prior distribution for the standard deviation of the random interviewer effects using (3.3), where the standard deviation s is given by the estimated standard deviation of the random interviewer effects for the same state in 2011, and used the weak priors on μ and Σ defined in Section 2.3. We implemented the Bayesian approach using PROC MCMC in the SAS software, and annotated examples of the code used are available in the supplemental materials.

5.3 Results

Figure 5.1 presents four scatter plots, enabling comparisons of the naïve estimates of the interviewer effects on the mean of perceived health status for each of the 50 states and the District of Columbia with the adjusted estimates based on the anchoring method, the "standard" adjustment method, and the two alternative Bayesian approaches to implementing the anchoring method. All estimates of interviewer effects were computed using (5.3).

The plots vary in terms of the methods used to implement the estimation approaches. We first consider a plot of the adjusted estimates of the interviewer effects based on the anchoring method against the naïve estimates of the interviewer effects from (5.3), using the frequentist approach described above (Figure 5.1a). The next plot (Figure 5.1b) presents the adjusted estimates based on the "standard" adjustment approach of including the covariates in a multilevel model. The third plot (Figure 5.1c) considers the first Bayesian anchoring method with a non-informative prior. Finally, the fourth plot (Figure 5.1d) once again considers the Bayesian anchoring method, only this time with the aforementioned informative prior based on analyses of the 2011 BRFSS data.

In general, we see that the anchoring method has a tendency to reduce estimates of the interviewer effects, regardless of the approach used. Data points below the 45-degree lines in each plot indicate states where a particular adjustment method reduced the estimates of the interviewer effects. In particular, the "standard" adjustment method will more often *increase* estimates of the interviewer effects in a non-trivial fashion relative to the naïve approach (Figure 5.1b).

Table 5.1 presents mean estimates and ranges of the interviewer effects across the 50 states and D.C. under the different methods. The anchoring method tended to reduce the estimates relative to the naïve method more often than the adjustment method, with 88.2% and 72.5% of states seeing a reduction in the estimated interviewer effects when using the frequentist and informative Bayesian anchoring methods, respectively (compared to only 60.8% of states when using the adjustment method). There is evidence in Table 5.1 that the use of prior information helps when applying the Bayesian anchoring method, but the frequentist version of the anchoring method still has the best performance overall. In some cases these reductions in the interviewer effect relative to the naïve approach were substantial: five of the states had reductions in the estimated interviewer effect of at least 33% regardless of the type of anchoring method used. In some cases, the anchoring approach did lead to slight increases in the estimated interviewer effects. These were predominantly cases where the interviewer effects were very small (suggesting that

the proposed adjustment would not be necessary, and that any resulting increases in the estimates were simply noise).

Figure 5.1 Scatter plots comparing the anchoring and naïve estimates of the interviewer effects for the 50 states and the District of Columbia, by estimation approach (NI = non-informative prior; Inf = weakly informative prior, based on analyses of the 2011 BRFSS data). Points belowthe 45-degree lines in each plot indicate states where a particular adjustment method reduced the estimates of the interviewer effects below that of the naïve estimate.

Figure 5.1a: Frequentist - Anchoring

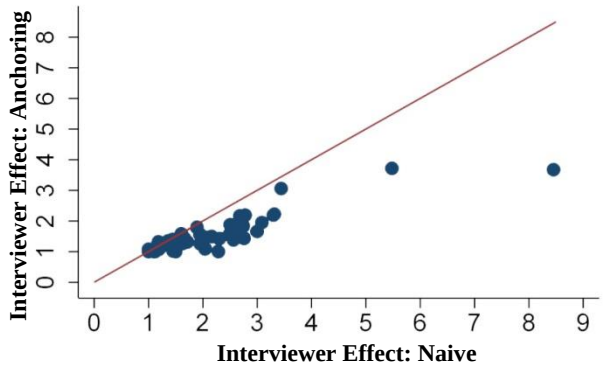


Figure 5.1b: Frequentist - Adjustment

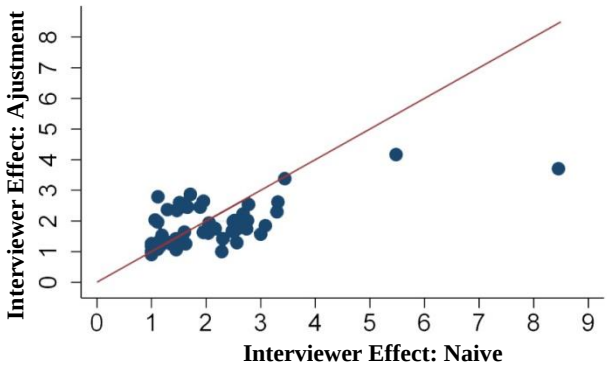


Figure 5.1c: Bayesian - Anchoring (NI)

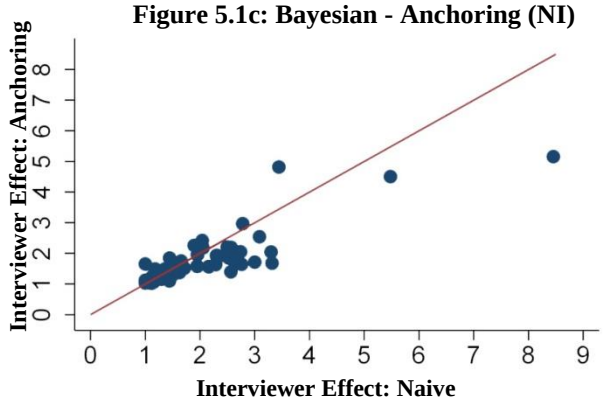


Figure 5.1d: Bayesian - Anchoring (Inf)

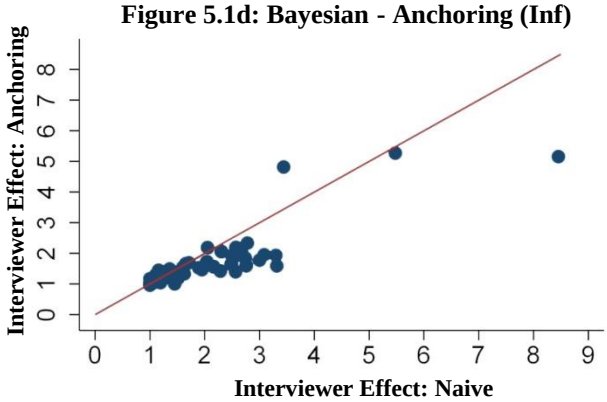


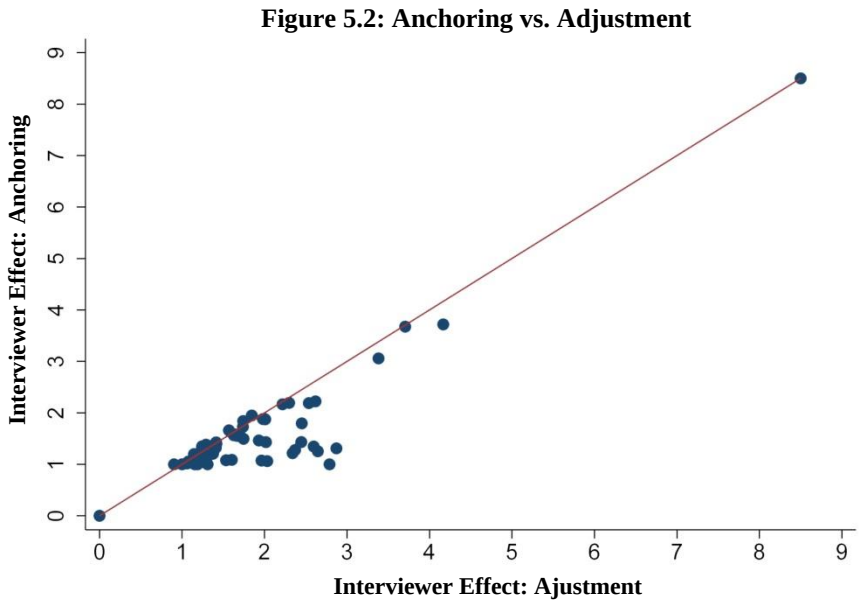
Table 5.1
Means and ranges of interviewer effects across the 50 States and the District of Columbia under the competing approaches

Estimation Approach	Interviewer Effects: Mean (Range)	Percentage of States with a Reduction
Frequentist – Naive	2.06 (1.00 – 8.45)	-
Frequentist – Adjustment	1.85 (0.90 – 4.17)	60.8%
Frequentist – Anchoring	1.51 (1.00 – 3.72)	88.2%
Bayesian – Anchoring, Non-Informative	1.79 (1.03 – 5.16)	58.8%
Bayesian – Anchoring, Informative	1.70 (0.96 – 5.27)	72.5%

When comparing the anchoring method with the "standard" adjustment method, we found consistent evidence of the anchoring method producing larger reductions in the estimated interviewer effects.

Figure 5.2 compares the estimated interviewer effects for the 50 states and D.C. when using the anchoring method and the adjustment method, considering the frequentist results only. We see that the interviewer estimates based on the adjustment method tend to be larger than the estimates based on the anchoring approach.

Figure 5.2 Scatter plot comparing the anchoring and adjusted estimates of the interviewer effects for the 50 states and the District of Columbia.



In general, we did not find significant benefits of using a Bayesian approach to implement the anchoring method in this application. We did find that for 92.5% of the states, the 95% credible interval for the interviewer variance component was smaller in width when using the informative prior than the credible interval based on the non-informative prior, as would be expected. However, the posterior medians of the interviewer variance components tended to be similar based on both Bayesian anchoring methods (Pearson correlation = 0.73).

6. Discussion

We have developed and evaluated a new method for estimating interviewer effects in the absence of interpenetrated assignment of sampled units to interviewers. Via a simulation study and applications using real survey data from the BRFSS, we have demonstrated the ability of the proposed anchoring method to improve estimates of interviewer effects in situations where interpenetrated assignment may not be feasible and interviewer variance may be arising from the underlying sample assignments. The anchoring method can also easily be applied in a Bayesian framework, leveraging prior information to improve the quality of predictions and inferences related to interviewer components of variance.

In interviewer-administered survey data collections, interviewer effects should generally be monitored as part of an ongoing data collection to prevent excessive problems with interviewer variance in survey

outcomes at the conclusion of the data collection. Survey managers responsible for this type of monitoring will likely benefit from the anchoring method, improving any real-time intervention decisions made for individual interviewers in a responsive survey design framework. Real-time interventions/re-trainings for interviewers who are found to have extreme effects on production outcomes or variables of scientific interest that in reality only reflect the features of the areas in which they are working and not actual interviewer performance will be at best inefficient and at worst could cause interviewers who are otherwise performing well to be inappropriately criticized and perhaps to leave a given study.

When using the anchoring method in practice, we would suggest that it be described as a method that "adjusts estimates of interviewer variance components for spurious within-interviewer correlation in survey measures of interest that can arise due to non-random assignment of sampled units to interviewers." We emphasize the importance of a sound theoretical selection of an anchoring variable (or variables) that ideally has the optimal properties described in this paper. In the absence of an anchoring variable with these optimal properties, we argue that "clean" estimation of interviewer variance in a non-interpenetrated sample design may simply not be possible, and that analysts 1) adjust for as many respondent-, interviewer-, and area-level covariates as possible when attempting to estimate the interviewer variance, and 2) report estimates of uncertainty associated with the estimated variance components, preferably using Bayesian approaches. This will prevent over-estimation of interviewer variance components and possible attribution of lower data quality to interviewers that are already performing extremely challenging tasks in the field.

There are several limitations to our proposed method. Perhaps the largest is the requirement for "anchoring" variables to not be subject to interviewer error and still be highly correlated with the substantive variable of interest. In our BRFSS example, we considered age, gender, race, and education as "anchoring" self-rated health. Although age is self-reported and thus possibly subject to some degree of measurement error (for example, reporting younger ages or rounding ages), we do not see an obvious mechanism by which this would be induced by the interviewer, although of course that possibility remains. A similar argument can be made for the other three factors, although the possibility of interviewer induced measurement error is slightly stronger due to issues such as "liking" between interviewers and respondents (West and Blom, 2017). In addition, the normality assumption that we make in the paper is highly restrictive. To deal with this in our application, we replaced the multivariate anchoring model (3.2) with a model that summarized the multiple anchors into a linear predictor that we then used in the bivariate anchoring model (3.1). While this linear predictor is effectively a sufficient statistic in the case where all of the anchoring variables are normal, as shown in the simulation study, it is more of an ad-hoc solution when some or all of the anchoring variables are non-normal, as was the setting in our application.

A more principled solution when one or more of the components of Y are dichotomous variable would be to consider extensions such as probit random effects models, replacing y_{ijk} in (3.1) with a latent y_{ijk}^*, where the observed $y_{ijk} = I\left(y_{ijk}^* > 0\right)$ and the variance $\sigma_k^2 = 1$ for identifiability, for all values of k where

y_{ijk} is dichotomous. More ambitiously, we could use a Gaussian random effects copula model (Wu and de Leon, 2014) for arbitrary distributions for Y^*. Standard software will not accommodate such models, although methods that integrate over random effects or use fully Bayesian approaches could be considered. Next, while other sources of measurement error are potentially important to address in inference, our focus here is on measurement error variance introduced by interviewer effects and its estimation in the absence of interpenetration. Finally, we note that our approach, like its competitors, relies on observed data and is thus not a replacement for true interpenetration, which ensures that all forms of non-random assignment (observed and unobserved) are eliminated.

In addition to extending the anchoring method to the case of regression coefficients and non-normal variables, future applications also need to consider contexts where the correlations of the anchoring variables with the survey variables of interest that may be prone to interviewer effects are modest at best. Our simulation study suggests that good anchoring variables having strong associations with the survey variables of interest are important for the effectiveness of this method, and future studies should also focus on the identification of sound anchoring variables (like age, education, etc.) that are unlikely to be affected by interviewers and could serve as useful anchors in other applications.

Acknowledgements

Funding for this research was provided by NIH Grant #1R01AG058599-01. The authors would like to thank the editor, associate editor, and two reviewers for their guidance which has improved the manuscript.

Supplemental materials

SAS code implementing the different approaches

The SAS code below can be used to implement the anchoring method using a standard frequentist approach. Implementing this approach requires the data to be in a “long” structure with two observations per subject (corresponding to the two variables), where the variable X2 is an indicator variable for the anchoring variable (1 = the observation on Y is the anchor, 0 = the observation on Y is the variable of interest), the variable X1 is an indicator for the variable of interest (1 = the observation on Y is the variable of interest, 0 = the observation on Y is the anchor), INTVID is the interviewer ID, and OBS is a subject ID:

```
proc mixed data=yourlongdata;
  class INTVID;
  model y = x2 / solution;
  random x1 / sub=INTVID;
  repeated / sub=obs type=un r rcorr;
run;
```

The SAS code below can be used to fit the naïve model using a Bayesian approach with a weakly informative prior. This approach requires the data to be in the same "long" format:

```
proc mcmc data=yourlongdata seed=41279 nmc=20000 thin=25;
  where x1 = 1; /* only fit model to variable of interest */
  parms B0 S2;
  parms Sigma 1;
  prior B: ~ normal(0, var=1e6); /* prior for means */
  prior S2 ~ igamma(0.01, scale = 0.01); /* NI prior for resid. var. */
  prior Sigma ~ t(0, sd=0.045, df=3, lower=0); /* informative prior for SD of
  interviewer effects, per Gelman (2006); SD of distribution is estimated SD
  of random interviewer effects from 2011, constrains posterior */
  random Gamma ~ normal(0,sd=Sigma) subject=INTVID;
  Mu = B0 + Gamma; /* model with interviewer effects only for variable of
  interest */
  model y ~ normal(Mu,var=S2);
run;
```

Finally, the SAS code below can be used to implement the anchoring method using a Bayesian approach with a weakly informative prior. Implementing this approach requires the data to be in a wide structure, with one row per case and interviewer IDs (INTVID):

```
proc mcmc data=yourwidedata seed=41279 nmc=20000 thin=25;
  array y[2] genhlthmdd age10; /* var1=variable of interest, var2=anchor */
  array Mu[2]; /* vector of two observations for each case */
  array Cov[2,2]; /* residual covariance matrix */
  array S[2,2]; /* for defining prior of COV */
  array H[2] 0 H1; /* H1 = fixed effect for change in mean for anchor */
  parms B0 Cov; /* intercept (mean of variable of interest) and residual
  covariance matrix */
  parms H1 0; /* change in mean for anchor */
  parms Sigma 1;
  prior B: H: ~ normal(0, var=1e6); /* normal prior for fixed effects */
  prior Cov ~ iwish(2,S); /* prior for 2x2 residual covariance matrix */
  prior Sigma ~ t(0, sd=0.045, df=3, lower=0); /* informative prior for SD of
  interviewer effects, per Gelman (2006); SD of distribution is estimated SD
  of random interviewer effects from 2011, constrains posterior */
  begincnst;
    call identity(S); /* use identity matrix in defining prior for residual
    covariance matrix (non-informative) */
  endcnst;
  random Gamma ~ normal(0,sd=Sigma) subject=INTVID;
  Mu[1] = B0 + Gamma; /* interviewer effect only applies to variable of
  interest */
  Mu[2] = B0 + H1; /* mean for anchor (note: this parameterization used to
  ensure easy calculation of posterior SD of B0 for interviewer effects */
  model y ~ mvn(Mu, Cov);
run;
```

References

Biemer, P.P. (2010). Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly*, 74(5), 817-848.

Biemer, P.P., and Stokes, S.L. (1985). Optimal design of interviewer variance experiments in complex surveys. *Journal of the American Statistical Association*, 80(389), 158-166.

Brunton-Smith, I., Sturgis, P. and Williams, J. (2012). Is success in obtaining contact and cooperation correlated with the magnitude of the interviewer variance? *Public Opinion Quarterly*, 76, 265-286.

Carle, A.C. (2009). Fitting multilevel models in complex survey data with design weights: Recommendations. *BMC Medical Research Methodology*, 9, 49-62.

Centers for Disease Control (2013). Behavioral Risk Factor Surveillance System: OVERVIEW: BRFSS 2012. Accessed at http://www.cdc.gov/brfss/annual_data/2012/pdf/Overview_2012.pdf.

Cernat, A., and Sakshaug, J.W. (2021). Interviewer effects in biosocial survey measurements. *Field Methods*, 33, 236-252.

Durrant, G.B., Groves, R.M., Staetsky, L. and Steele, F. (2010). Effects of interviewer attitudes and behaviors on refusal in household surveys. *Public Opinion Quarterly*, 74, 1-36.

Elliott, M.R., and West, B.T. (2015). "Clustering by interviewer": A source of variance that is unaccounted for in single-stage health surveys. *American Journal of Epidemiology*, 182, 118-126.

Fellegi, I.P. (1974). An improved method of estimating the correlated response variance. *Journal of the American Statistical Association*, 69,496-501.

Fowler, F.J., and Mangione, T.W. (1989). *Standardized Survey Interviewing: Minimizing Interviewer-Related Error*. Newbury Park: Sage.

Franks, P., Gold, M.R. and Fiscella, K. (2003). Sociodemographics, self-rated health, and mortality in the US. *Social Science & Medicine*, 56, 2505-2514.

Gao, S., and Smith, T.M.F. (1998). A constrained MINQU estimator of correlated response variance from unbalanced data in complex surveys. *Statistica Sinica*, 8, 1175-1188.

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

Groves, R.M. (2004). Chapter 8: The interviewer as a source of survey measurement error. *Survey Errors and Survey Costs (2nd Edition)*. New York: Wiley-Interscience.

Groves, R.M., and Magilavy, L.J. (1986). Measuring and explaining interviewer effects in centralized telephone surveys. *Public Opinion Quarterly*, 50, 251-266.

Heeringa, S.G., West, B.T. and Berglund, P.A. (2017). *Applied Survey Data Analysis, Second Edition*. Boca Raton, FL: Chapman Hall/CRC Press.

Hox, J.J. (1994). Hierarchical regression models for interviewer and respondent effects. *Sociological Methods and Research*, 22, 300-318.

Joffe, M.M., Ten Have, T.R., Feldman, H.I. and Kimmel, S.E. (2004). Model selection, confounder control, and marginal structural models: Review and new applications. *The American Statistician*, 58, 272-279.

Kalton, G. (1983). *Introduction to Survey Sampling*, Sage Publications: London, UK.

Kish, L. (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Association*, 57, 92-115.

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

Kleffe, J., Prasad, N.G.N. and Rao, J.N.K. (1991). "Optimal" estimation of correlated response variance under additive models. *Journal of the American Statistical Association*, 86, 144-150.

Laird, N.M., and Ware, J.H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38, 963-974.

Lepkowski, J.M., Mosher, W.D., Groves, R.M., West, B.T., Wagner, J. and Gu, H. (2013). Responsive design, weighting, and variance estimation in the 2006-2010 National Survey of Family Growth. National Center for Health Statistics. *Vital Health Stat*, 2(158).

Liang, K.-Y., and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.

O'Muircheartaigh, C.A., 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*, 161, 63-77.

Pfeffermann, D., Skinner, C.J., Holmes, D.J., Goldstein, H. and Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society, Series B*, 60, 23-40.

Rabe-Hesketh, S., and Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society-A*, 169, 805-827.

Rasbash, J., and Goldstein, H. (1994). Efficient analysis of mixed hierarchical and cross-classified random structures using a multilevel model. *Journal of Educational and Behavioral Statistics*, 19, 337-350.

Rohm, T., Carstensen, C.H., Fischer, L. and Gnambs, T. (2021). Disentangling interviewer and area effects in large-scale educational assessments using cross-classified multilevel item response models. *Journal of Survey Statistics and Methodology*, 9, 722-744.

Sakshaug, J.W., Tutz, V. and Kreuter, F. (2013). Placement, wording, and interviews: identifying correlates of consent to link survey and administrative data. *Survey Research Methods*, 7, 133-144.

Schaeffer, N.C., Dykema, J. and Maynard, D.W. (2010). Interviewers and Interviewing. In *Handbook of Survey Research, Second Edition* (Eds., J.D. Wright and P.V. Marsden), Bingley, U.K.: Emerald Group Publishing Limited.

Schnell, R., and Kreuter, F. (2005). Separating interviewer and sampling-point effects. *Journal of Official Statistics*, 21, 389-410.

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

Stiratelli, R., Laird, N. and Ware, J. (1984). Random effects models for serial observations with binary responses. *Biometrics*, 40, 961-971.

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*, 180, 531-550.

Veiga, A., Smith, P.W.F. and Brown, J.J. (2014). The use of sample weights in multivariate multilevel models with an application to income data collected by using a rotating panel survey. *Journal of the Royal Statistical Society (Series C)*, 63, 65-84.

von Sanden, N., and Steel, D. (2008). Optimal estimation of interviewer effects for binary response variables through partial interpenetration. Centre for Statistical and Survey Methodology, University of Wollongong, Working Paper 04-08.

West, B.T., and Blom, A.G. (2017). Explaining interviewer effects: A research synthesis. *Journal of Survey Statistics and Methodology*, 5, 175-211.

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.

West, B.T., and Olson, K. (2010). How much of interviewer variance is really nonresponse error variance? *Public Opinion Quarterly*, 74, 1004-1026.

West, B.T., Kreuter, F. and Jaenichen, U. (2013). "Interviewer" effects in face-to-face surveys: A function of sampling, measurement error or nonresponse? *Journal of Official Statistics*, 29, 277-297.

Wu, B., and de Leon, A.R. (2014). Gaussian copula mixed models for clustered mixed outcomes, with application in developmental toxicology. *Journal of Agricultural, Biological, and Environmental Statistics*, 19, 39-56.

Survey Methodology, June 2022
Vol. 48, No. 1, pp. 49-71
Statistics Canada, Catalogue No. 12-001-X

Relative performance of methods based on model-assisted survey regression estimation: A simulation study

Erin R. Lundy and J.N.K. Rao[1]

Abstract

Use of auxiliary data to improve the efficiency of estimators of totals and means through model-assisted survey regression estimation has received considerable attention in recent years. Generalized regression (GREG) estimators, based on a working linear regression model, are currently used in establishment surveys at Statistics Canada and several other statistical agencies. GREG estimators use common survey weights for all study variables and calibrate to known population totals of auxiliary variables. Increasingly, many auxiliary variables are available, some of which may be extraneous. This leads to unstable GREG weights when all the available auxiliary variables, including interactions among categorical variables, are used in the working linear regression model. On the other hand, new machine learning methods, such as regression trees and lasso, automatically select significant auxiliary variables and lead to stable nonnegative weights and possible efficiency gains over GREG. In this paper, a simulation study, based on a real business survey sample data set treated as the target population, is conducted to study the relative performance of GREG, regression trees and lasso in terms of efficiency of the estimators and properties of associated regression weights. Both probability sampling and non-probability sampling scenarios are studied.

Key Words: Model assisted inference, Calibration estimation; Model selection; Generalized regression estimator.

1. Introduction

At Statistics Canada and several other statistical agencies, there is a growing interest in leveraging auxiliary data, possibly from administrative sources, to improve the efficiency of estimators. Machine learning techniques have become a popular tool in various disciplines for utilizing such auxiliary information. These methods often do not require the distributional assumptions of more traditional methods and are able to adapt to complex non-linear and non-additive relationships between the outcomes and auxiliary variables. Machine learning methods have been applied to survey data in a variety of contexts such as response/adaptive designs, data processing, nonresponse adjustment and weighting (Buskirk, Kirchner, Eck and Signorino, 2018; Kern, Klausch and Kreuter, 2019).

Recently, the use of machine learning techniques to improve the efficiency of estimators of totals and means through model-assisted survey regression estimation under probability sampling has been considered. Model-assisted survey regression estimators of finite population totals may reduce variability and lead to significant gains in efficiency if the available auxiliary variables are strongly associated with the survey variable of interest. Increasingly, a large number of auxiliary variables are available, some of which may be extraneous. In this case, variable selection followed by regression estimation based on the selected model may improve efficiency of the survey regression estimators of finite population totals. We consider finite population estimation using the generalized regression (GREG) estimator with various linear working models (Särndal, Swensson and Wretman, 1992). Model-assisted estimators, using lasso

1. Erin R. Lundy, Statistical Integration Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6. E-mail: erin.lundy@statcan.gc.ca; J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, K1S 5B6.

and adaptive lasso methods (McConville, Breidt, Lee and Moisen, 2017) and regression trees (McConville and Toth, 2019), have been applied to survey data. Other nonlinear models, such as penalized splines and neural networks, have been explored for model-assisted estimation; see Breidt and Opsomer (2017) for a survey of these techniques.

Another field of research where the use of model-assisted estimators has been proposed is estimation from non-probability samples. Increasing costs and declining response rates are leading to an expanding interest in the use of non-probability samples. However, the process generating a non-probability sample is unknown and such samples are subject to selection bias. Two commonly used approaches to estimation from non-probability samples are quasi-randomization and superpopulation modeling. In the first, the sample is treated as if it was obtained from probability sampling but with unknown selection probabilities. The pseudo-inclusion probabilities are estimated via a propensity model that uses the sample data in combination with some external data set that covers the targeted population. Machine learning techniques have been employed in the estimation of pseudo-inclusion probabilities or, equivalently, in the construction of pseudo-weights. Kern, Li and Wang (2020) investigated several machine learning techniques to construct pseudo-weights using a propensity score-based kernel weighting for non-probability samples. Rafei, Flannagan and Elliott (2020) developed a pseudo-weighting approach using Bayesian Additive Regression Trees.

In the superpopulation approach, observed values of the variables of interest are assumed to be generated by some model. The model is estimated from the data and, along with external population control data, is used to project the sample to the population. Under this framework, calibration to known population totals of auxiliary variables provides a means of potentially reducing the effect of sample selection bias. Chen, Valliant and Elliott (2018) discussed the implementation of model calibration using adaptive lasso for data based on non-probability sampling. In scenarios where the population totals are estimated, Chen, Valliant and Elliott (2019), incorporated the sampling uncertainty of the benchmarked data, obtained from a probability sample survey, into the variance component of a model-assisted calibration estimator using adaptive lasso regression. Therefore, unlike in the probability sampling context where the use of model-assisted estimation seeks to improve the efficiency of estimators, the use of these techniques in a non-probability sampling context aims to diminish the impact of selection bias.

We consider several lasso-based estimators as well as a regression tree estimator and evaluate their performance in both a probability sampling context and a non-probability sampling set up. In Section 2, the model-assisted estimators considered are discussed. The set up for a simulation study under probability sampling is described in Section 3. The results of the simulation study on the root mean square error of the estimators, relative bias of variance estimators and properties of survey weights are presented in Section 4. Except for the GREG estimator, all the model-assisted estimators considered here involve variable selection and yield, if applicable, regression weights that depend on the survey variable of interest, y. The impact of using a single set of regression weights for multiple related study variables is also investigated in this section. The results of the simulation study using a non-probability sampling scenario are detailed in Section 5. We conclude with a summary of the findings in Section 6.

2. Model-assisted estimation under probability sampling

2.1 GREG estimators

Consider the estimation of a finite population total $t_y = \sum_{i \in U} y_i$, where $U = \{1, \ldots, N\}$ is the set of units of the finite population and y_i is the value of the survey variable of interest for the unit $i \in U$. Let $s \subset U$ be a sample selected according to a sampling design $p(.)$, where $p(s)$ is the probability of selecting s. For $i \in U$, let $\pi_i = \Pr[i \in s]$ denote the first-order inclusion probabilities of the design. We assume $\pi_i > 0$ for all $i \in U$. Additionally, assume d auxiliary variables, $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{id})^T$ are known for each $i \in U$. A standard approach is to use the Horvitz-Thompson estimator

$$\hat{t}_{y,\mathrm{HT}} = \sum_{i \in s} \frac{y_i}{\pi_i} = \sum_{i \in s} d_i y_i$$

where $d_i = \pi_i^{-1}$ denotes design weights. Under this strictly design-based framework, the auxiliary data do not impact the form of the estimator but can impact the design weights, d_i, through the specification of the sampling design.

One strategy to use auxiliary data in estimation is to employ a model-assisted estimator of t_y by specifying a working model for the mean of y given $\mathbf{x}$ and use this model to predict y values. Specifying a linear regression working model leads to the generalized regression (GREG) estimator (Cassel, Särndal and Wretman, 1976). The GREG estimator typically has smaller variance than the Horvitz-Thompson estimator if the working model has some predictor power for y. Here, we consider the GREG estimator under a linear regression working model

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i \tag{2.1}$$

with $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$, ε_i independent and identically distributed with mean zero and variance σ^2 and $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ip})^T$. The GREG estimator is given by

$$\hat{t}_{y,\mathrm{GREG}} = \sum_{i \in s} \frac{y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_s}{\pi_i} + \sum_{i \in U} \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_s \tag{2.2}$$

with the regression coefficients $\boldsymbol{\beta}$ estimated as

$$\hat{\boldsymbol{\beta}}_s = \underset{\beta}{\operatorname{argmin}} \left(\mathbf{Y}_s - \mathbf{X}_s \boldsymbol{\beta}\right)^T \boldsymbol{\Pi}_s^{-1} \left(\mathbf{Y}_s - \mathbf{X}_s \boldsymbol{\beta}\right) = \left(\mathbf{X}_s^T \boldsymbol{\Pi}_s^{-1} \mathbf{X}_s\right)^{-1} \mathbf{X}_s^T \boldsymbol{\Pi}_s^{-1} \mathbf{Y}_s, \tag{2.3}$$

where $\mathbf{X}_s$ is a $n \times (p+1)$ matrix, Y_s is a n-vector and Π_s is an $n \times n$ diagonal matrix of first-order inclusion probabilities for the sampled units.

The GREG estimator can also be written as a weighted sum of the variable of interest, y, yielding regression weights that are independent of y and, therefore, can be applied to any study variable, y:

$$\hat{t}_{y,\mathrm{GREG}} = \sum_{i \in s} \left[1 + \left(\mathbf{t}_x - \hat{\mathbf{t}}_{x,\mathrm{HT}}\right)^T \left(\sum_{k \in s} \mathbf{x}_k \mathbf{x}_k^T d_k \right)^{-1} \mathbf{x}_i \right] d_i y_i = \sum_{i \in s} w_i y_i \tag{2.4}$$

where $\mathbf{t}_x$ is the known population total vector of the covariates $\mathbf{x}$ and $\hat{\mathbf{t}}_{x,\mathrm{HT}}$ is the Horvitz-Thompson estimator vector of the covariate population totals $\mathbf{t}_x = \sum_{i \in U} \mathbf{x}_i$. The regression weights, w_i, are termed calibration weights because they satisfy the calibration constraint $\sum_{i \in s} w_i \mathbf{x}_i = \sum_{i \in U} \mathbf{x}_i$. The calibration weight w_i does not depend on the study variable y_i. Note that the GREG estimator (2.4) can alternatively be expressed as

$$\hat{t}_{y,\mathrm{GREG}} = \hat{t}_{y,\mathrm{HT}} + \left(\mathbf{t}_x - \hat{\mathbf{t}}_{x,\mathrm{HT}}\right)^T \hat{\boldsymbol{\beta}}_s$$

which only requires known population totals $\mathbf{t}_x$. For the GREG estimator, the individual population values $\mathbf{x}_i$, $i \in U$ are not needed.

If a variable selection procedure, such as a forward stepwise procedure, is implemented prior to fitting the linear regression model, then the calibration weights will depend on y as the selected models may vary across study variables. This type of stepwise survey regression estimator is calibrated to the auxiliary variables selected by the variable selection procedure for a specific variable of interest, y.

Using a working linear regression model with many auxiliary variables, including interactions of categorical auxiliary variables, can produce substantially variable weights, and greatly increase the variance of the GREG estimator. Furthermore, some of the regression weights, w_i, $i \in s$, may be negative, thus losing the interpretation of a weight as the number of population units represented by the sampled unit.

2.2 Survey regression estimator with Lasso

If the linear regression model in (2.1) is sparse, i.e., p is large, and, say, only p_0 of the p regression coefficients are nonzero, then the estimation of the zero coefficients in (2.3) leads to extra variation in the GREG estimator (2.2). In this case, model selection to remove extraneous variables could reduce the overall design variance of the GREG estimator, leading to more efficient estimates of finite population totals. The least absolute shrinkage and selection operator (lasso) method, developed by Tibshirani (1996), simultaneously performs model selection and coefficient estimation by shrinking some regression coefficients to zero. The lasso approach estimates coefficients by minimizing the sum of squared residuals subject to a penalty constraint on the sum of the absolute value of the regression coefficients.

McConville et al. (2017) proposed using survey-weight lasso estimated regression coefficients given by

$$\hat{\boldsymbol{\beta}}_{s,L} = \underset{\beta}{\operatorname{argmin}} \left(\mathbf{Y}_s - \mathbf{X}_s \boldsymbol{\beta}\right)^T \boldsymbol{\Pi}_s^{-1} \left(\mathbf{Y}_s - \mathbf{X}_s \boldsymbol{\beta}\right) + \lambda \sum_{j=1}^{p} \left|\beta_j\right|,$$

where $\lambda \geq 0$. The lasso survey regression estimator for the total t_y is then given by

$$\hat{t}_{y,\mathrm{LASSO}} = \sum_{i \in s} \frac{y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{s,L}}{\pi_i} + \sum_{i \in U} \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{s,L}.$$

The value of the penalty parameter λ must be selected prior to obtaining the estimated coefficients. In general, this process of specifying hyperparameters prior to fitting the final model is called hyperparameter tuning. There are several potential selection criteria that can used to select the value of hyperparameters including Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) or cross-validation. We used a version of cross-validation which incorporates the design weights in our simulation study; see McConville (2011) for discussion of the selection of the penalty parameter for survey-weighted lasso coefficient estimates.

2.3 Survey regression estimator with adaptive Lasso

An issue with the use of the lasso criterion is that by shrinking the regression coefficients towards zero it yields biased estimates for regression coefficients that are far from zero. Under the adaptive lasso criterion (Zou, 2006), the coefficients in the l_1 penalty are weighted by the inverse of a root-n consistent estimator of $\boldsymbol{\beta}$. Therefore, the bias for large coefficients tends to be smaller.

McConville et al. (2017) considered an adaptive lasso survey regression estimator

$$\hat{t}_{y,\,\mathrm{ALASSO}} = \sum_{i\in s} \frac{y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{s,\,\mathrm{AL}}}{\pi_i} + \sum_{i\in U} \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{s,\,\mathrm{AL}},$$

where

$$\hat{\boldsymbol{\beta}}_{s,\,\mathrm{AL}} = \underset{\beta}{\operatorname{argmin}} \left(\mathbf{Y}_s - \mathbf{X}_s\boldsymbol{\beta}\right)^T \boldsymbol{\Pi}_s^{-1} \left(\mathbf{Y}_s - \mathbf{X}_s\boldsymbol{\beta}\right) + \lambda \sum_{j=1}^{p} \frac{\left|\beta_j\right|}{\left|\hat{\beta}_{sj}\right|}$$

and $\hat{\boldsymbol{\beta}}_s$ is given by (2.3). The reliance of the adaptive lasso method on the standard weighted linear regression coefficient estimates, $\hat{\boldsymbol{\beta}}_s$, leads to a loss of efficiency in settings when p is large because the estimates $\hat{\boldsymbol{\beta}}_s$ tend to be very unstable.

2.4 Lasso calibration estimators

The lasso and adaptive lasso methods do not produce regression weights directly, as the estimators cannot be expressed as weighted combinations of the y-values. McConville et al. (2017) developed lasso survey regression weights using a model calibration approach and a ridge regression approximation. These lasso regression weights depend on the variable of interest, y.

The lasso calibration estimator is calculated by regressing the variable of interest, y_i, on an intercept and the lasso-fitted mean function $\mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{s,\,L}$. The lasso calibration estimator can be written in the same form as (2.4), where $\mathbf{x}_i$ is replaced by $\mathbf{x}_i^* = \left(1, \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{s,\,L}\right)^T$:

$$\hat{t}_{y,\,\mathrm{CLASSO}} = \sum_{i\in s} \left[1 + \left(\mathbf{t}_{x^*} - \hat{\mathbf{t}}_{x^*,\,\mathrm{HT}}\right)^T \left(\sum_{k\in s} \mathbf{x}_k^* \mathbf{x}_k^{*T} d_k \right)^{-1} \mathbf{x}_i^* \right] d_i y_i. \tag{2.5}$$

Similarly, the adaptive lasso calibration estimator is given by

$$\hat{t}_{y,\text{CALASSO}} = \sum_{i\in s}\left[1+\left(\mathbf{t}_{x^{**}} - \hat{\mathbf{t}}_{x^{**},\text{HT}}\right)^T\left(\sum_{k\in s}\mathbf{x}_k^{**}\mathbf{x}_k^{**T}d_k\right)^{-1}\mathbf{x}_i^{**}\right]d_i y_i,$$

where the lasso-fitted mean for $\mathbf{x}_i^*$ in (2.5) is replaced by the adaptive lasso fit, $\mathbf{x}_i^{**} = \left(1, \mathbf{x}_i^T\hat{\boldsymbol{\beta}}_{s,\text{AL}}\right)^T$. The weights for the lasso calibration estimators are calibrated to the population size N and to the population total of the lasso-fitted mean functions.

2.5 Regression tree estimator

The GREG estimator can also be expressed as

$$\hat{t}_{y,r} = \sum_{i\in s}\frac{y_i - \hat{h}_n(\mathbf{x}_i)}{\pi_i} + \sum_{i\in U}\hat{h}_n(\mathbf{x}_i), \tag{2.6}$$

where $\hat{h}_n(\mathbf{x}_i)$ is an estimator of the mean function of Y_i given $\mathbf{X}_i = \mathbf{x}_i$, $h(\mathbf{x}_i) = E\left(Y_i \mid \mathbf{X}_i = \mathbf{x}_i\right)$, based on the sample data $(y_i, \mathbf{x}_i), i \in s.$ As an alternative to a linear regression model, McConville and Toth (2019) proposed estimating $h(\mathbf{x})$ with a regression tree model using the following algorithm:

1. Let $k(n)$ be the minimum box size and α be a specified significance level.
2. If the dataset contains at least $2k(n)$ observations then continue to step 3; otherwise, stop.
3. Among the auxiliary variables $x_l, l = 1, \ldots, d,$ choose a variable to split the data. The chosen x_l is the variable that shows the largest significance difference after testing the null-hypothesis of homogeneous $E\left[y \mid x_l\right]$. If no variable leads to a significant difference, then stop.
4. Split the data into two sets S_L and S_R by splitting based on the value of the selected variable x_l that results in the largest decrease in the estimated mean square error, while satisfying the requirement that each subset contains at least $k(n)$ units.
5. For each of the resulting subsets of the data, return to step 1.

The resulting regression tree model groups the categories of an auxiliary variable based on their relationship to the variable of interest and only includes auxiliary variables and interactions associated with this variable. Importantly, including a categorical variable does not require a split for each category, potentially reducing the model size substantially while still capturing important interactions.

After fitting a regression tree model, we obtain a set of boxes $Q_n = \left\{B_{n1}, B_{n2}, \ldots, B_{nq}\right\}$ which partition the data. Let $I(\mathbf{x}_i \in B_{nk}) = 1$ if $\mathbf{x}_i \in B_{nk}$ and 0 otherwise, for $k = 1, \ldots, q.$ This means that $I(\mathbf{x}_i \in B_{nk}) = 1$ for exactly one box $B_{nk} \in Q_n$ for every $i \in s.$ For every $\mathbf{x}_i \in B_{nk}$, the estimator of $h(\mathbf{x}_i)$ is given by

$$\tilde{h}_n(\mathbf{x}_i) = \tilde{\#}(B_{nk})^{-1}\sum_{i\in s}\pi_i^{-1}y_i I(\mathbf{x}_i \in B_{nk}) = \tilde{\mu}_{nk}, \tag{2.7}$$

where

$$\tilde{\#}\left(B_{nk}\right)=\sum_{i\in s}\pi_i^{-1}I\left(\mathbf{x}_i\in B_{nk}\right)$$

is the HT estimator of the population size in box B_{nk}. The regression tree estimator $\hat{t}_{y,\text{TREE}}$ is obtained by inserting equation (2.7) into the generalized regression estimator, given in equation (2.6), leading to the post stratified estimator

$$\hat{t}_{y,\text{TREE}}=\sum_k N_k\tilde{\mu}_{nk},$$

where N_k is the number of units in U that belong to box k.

Since $\tilde{h}_n\left(\mathbf{x}_i\right)$ can be written as a linear regression estimator with q indicator function covariates, the regression tree estimator is also a post-stratified estimator, where each box B_{nk} represents a post-stratum. This implies that this estimator is calibrated to the population total of each box, providing a data-driven mechanism, dependent on y, for selecting post-strata that ensures that none of them are empty. As a result, the regression weights are guaranteed to be non-negative. The weights produced by this estimation procedure depend on the variable of interest, y. Therefore, unlike the GREG approach, a single set of generic weights to apply to all study variables is not available. Instead, a set of weights for each survey variable of interest is produced.

2.6 Variance estimation under stratified simple random sampling

Under stratified simple random sampling, a variance estimator of the model-assisted survey regression estimators described above is obtained by the Taylor linearization method and given by

$$\hat{V}\left(\hat{t}_y\right)=\sum_h\frac{N_h\left(N_h-n_h\right)}{n_h}\frac{1}{n_h-1}\sum_{i\in s_h}\left(e_{hi}-\bar{e}_h\right)^2, \tag{2.8}$$

where h indexes the strata, N_h is the number of population units in stratum h, n_h is the number of sampled units s_h in stratum h, $e_{hi}=y_{hi}-\hat{h}_n\left(\mathbf{x}_{hi}\right)$ is the residual of sample unit i in stratum h under the regression model and $\bar{e}_h$ is the average residual in stratum h.

The variance estimators readily extend to more complex sampling designs, but for simplicity we have given the expression only for stratified simple random sampling which is used in the simulation study of Section 3.

3. Simulation study using Financing and Growth of Small and Medium Enterprises Survey data

In this section, we describe a simulation study used to compare the performance of model-assisted survey regression estimators relative to the purely design-based HT estimator. Using the Survey of Financing and Growth of Small and Medium Enterprises data as the population, we compare the

estimators in repeated samples of the data to produce estimates of the total amount requested for trade credit which is a particular type of financing.

3.1 Simulation population

The Survey of Financing and Growth of Small and Medium Enterprises (SFGSME) is a periodic survey of enterprises which occurs approximately every three years and collects information on the types of financing businesses use. The sample is stratified by size, defined by the number of employees, the age of the business, industry at the 2-digit North American Industry Classification System (NAICS) and geography. A sample of approximately 17,000 enterprises was selected for the 2017 iteration of the survey.

The Business Register (BR) is the primary source of auxiliary information for business surveys at Statistics Canada. The frame used by the SFGSME was constructed by selecting from Statistic's Canada BR all enterprises with between 1 and 499 employees and a minimum gross revenue of $30,000. Non-profit enterprises as well as enterprises belonging to certain industry subgroups were excluded from the target population. The BR contains information on the location, number of employees, industry as well as revenue for each enterprise in the population.

3.2 Simulation methodology

We conducted a simulation study to compare the relative performance of several model-assisted survey regression estimators, using three and four categorical auxiliary variables. We considered sample sizes of $n = \{200; 500; 1{,}000\}$ from the 9,115 respondents in the SFGSME dataset. This dataset was treated as the target population and repeated samples were drawn using stratified simple random sampling as this is the design commonly used by statistical agencies for business surveys. We assumed there are two strata, where stratum A consists of units with revenue of less than $2.5 million and stratum B consists of units with revenue greater than $2.5 million. We assumed equal sample sizes in each stratum but most of the units in the population, approximately 70%, belong to stratum A. Under this sampling design, larger revenue units are over-represented, resulting in an unequal probability sampling design. Preliminary simulations using a simple random sample design were also considered and yielded similar results. The minimum sample size considered was $n = 200$ because for smaller sample sizes and 28 categories of x-variables, there were often categories without a sampled unit. In this case, it is not possible to calibrate the GREG estimator to all the pre-specified marginal totals.

For each sample, models using three x-variables, industry (10 categories), employment size (4 categories) and region (6 categories) were used to estimate total amount of trade credit requested and results were compared to the true total. We also considered a fourth variable, revenue, with 8 categories. For each combination of the three different sample sizes, and the two sets of auxiliary variables, with 20 and 28 main effects categories, we drew 5,000 repeated stratified random samples from the target

population. For each sample, we implemented the HT estimator and several model-assisted survey estimators as summarized in Table 3.1 below:

Table 3.1
Summary of model assisted estimators considered in simulation study

Estimator	Auxiliary Data	Regression Weights	Calibration Totals
GREG	Marginal totals Considered main effects only	Independent of y	All auxiliary variables
GREG with forward variable selection (FSTEP)	Individual values Considered main effects only	Dependent on y	Selected auxiliary variables
Regression Tree (TREE)	Individual values	Dependent on y, strictly positive	Population size of each box
Lasso (LASSO)	Individual values Considered main effects (1-way) and two-way interactions (2-way)		
Calibrated lasso (CLASSO)	Individual values Considered main effects (1-way) and two-way interactions (2-way)	Dependent on y	Population size and lasso-fitted mean function
Adaptive lasso (ALASSO)	Individual values Considered main effects only		
Calibrated adaptive lasso (CALASSO)	Individual values Considered main effects only	Dependent on y	Population size and lasso-fitted mean function

We initially also considered adaptive lasso and adaptive lasso calibration estimators using all main effects and 2-way interactions, but estimates of the coefficients under the GREG linear model, $\hat{\boldsymbol{\beta}}_s$, were highly unstable leading to singularity issues.

All computations were completed in R (Version 3.4.0, 2017). The HT, GREG, regression tree and lasso estimators were calculated using the package **mase** (McConville, Tang, Zhu, Li, Cheung and Toth, 2018) and the adaptive lasso coefficients were computed using the package **glmnet** (Friedman, Hastie, Simon, Qian and Tibshirani, 2017). The function `cv.glmnet` was used to select the value of the penalty parameter for the lasso estimators. We used a 10-fold cross validation procedure which allows for the inclusion of design weights. For the regression tree estimator, the minimum box size $k(n)$ was specified as 25 and the level of significance α was 0.05. We also considered a minimum box size of 10 units. For small sample sizes, there was a small gain in efficiency relative to a minimum box size of 25. For sample sizes of $n = 1{,}000$, different choices for the minimum box size yielded similar results in term of mean square error. Forward stepwise selection for the FSTEP estimator was based on minimizing the Akaike Information Criteria (AIC) and was performed using the function `stepAIC` in the MASS package (Ripley, Venables, Bates, Hornik, Gebhardt and Firth, 2017).

In regressing the amount of trade credit requested for the entire finite population on the 28 marginal categories, the adjusted coefficient of determination was approximately $R^2 = 0.22$ when both main effects and two-way interaction effects were considered. For the population model with main effects only the number of significant effects was 15 and for the population model with main effects and two-way interactions, there were 2 significant main effects and 29 significant interaction effects. These population-level results indicate that useful predictive models should be sparse and that there may be important two-way interactions.

Fitting regression tree models to the amount of trade credit requested resulted in 25 splits. The first split was based on revenue, indicating that this is the auxiliary data that is most strongly related to the amount of trade credit requested. There were splits based on all four of the auxiliary variables considered: revenue, industry, employment size and geography. This is consistent with the conclusions that useful predictive models should be sparse but allow for higher order interactions.

4. Results of the simulation study

4.1 Performance of estimators in terms of design MSE

We computed design bias and design mean square error (MSE) from the 5,000 total estimates by sample size and number of marginal categories. The percentage absolute relative design bias was less than 2 percent for all the estimators for all scenarios. As expected, for all estimators, the bias decreases as the sample size increases.

Figure 4.1 displays the MSE of the HT, GREG, GREG with forward variable selection, regression tree and calibrated lasso estimators by sample size, based on the 5,000 simulated samples. The MSE values are similar for the adaptive and non-calibrated versions of the lasso estimators. For all the estimators, the decrease in MSE is much more pronounced from $n = 200$ to $n = 500$ than from $n = 500$ to $n = 1{,}000$. This is likely due to the small sample size, relative to the number of categories for the auxiliary variables. It may not be possible to explore all the potential effects, particularly higher order effects, with only 200 sampled units.

Table 4.1 displays the ratio the design MSE of each estimator to the MSE of the HT estimator for the total amount of trade credit requested. For $n = 200$, the regression tree estimator and the lasso (2-way) estimator with two factor interaction effects are the only model-assisted estimators that provide any efficiency gains, relative to the HT estimator, when the number of categories of auxiliary variables used is large. As the sample size increases, the gains in efficiency of the model-assisted survey regression estimators, relative to the HT estimator, are essentially equal. Using any of the model-assisted estimators when $n = 1{,}000$ results in a slight gain in efficiency, relative the HT estimator. There is little efficiency advantage for model-assisted estimators over the HT estimator, indicating that the auxiliary variables are not strongly related to the variable of interest.

Figure 4.1 Comparison of mean square error for HT, GREG, FSTEP, regression tree and calibrated lasso estimators (1-way and 2-way) for the total amount of trade credit requested.

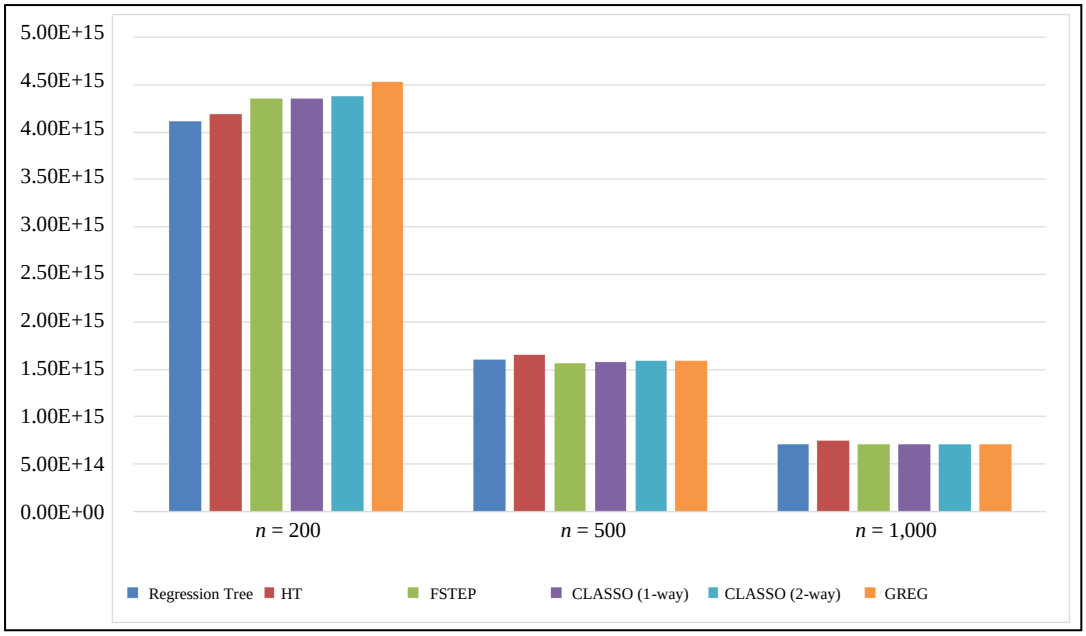


Table 4.1
Ratio of MSE of each estimator to MSE of HT estimator with 20 and 28 marginal categories

	20 categories			28 categories		
	$n=200$	$n=500$	$n=1{,}000$	$n=200$	$n=500$	$n=1{,}000$
GREG	1.067	1.011	0.994	1.084	0.959	0.954
FSTEP	1.036	1.009	0.994	1.040	0.945	0.958
TREE	1.023	1.007	0.977	0.983	0.963	0.949
LASSO (1-way)	1.020	0.995	0.986	1.009	0.946	0.947
CLASSO (1-way)	1.047	1.004	0.990	1.042	0.952	0.949
LASSO (2-way)	0.999	0.995	0.952	0.981	0.935	0.936
CLASSO (2-way)	1.061	1.029	0.966	1.045	0.959	0.950
ALASSO	1.024	0.999	0.986	1.021	0.948	0.948
CALASSO	1.040	1.005	0.989	1.037	0.951	0.949

The potential gains in efficiency for model-assisted estimators depend on the predictive power of the working model. In our simulation population, the strength of the relationship between the variable of interest and the available auxiliary variables is weak, leading to only slight efficiency gains relative to the purely design-based HT estimator. Therefore, to further explore the differences between the various model-assisted survey estimators, we ran additional simulations using different survey variables of interest, generated according to the following procedure:

1. Assuming a lasso model with main effects only, we obtained the lasso coefficient estimates for the amount of trade credit requested, y_i, using the population values for the auxiliary variables $\mathbf{x}_i$, including revenue.

2. We used the coefficient estimates $\hat{\beta}_L$ obtained in step 1 and the population values for $\mathbf{x}_i$ to generate a new survey variable of interest

$$y_i^* = \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_L + \mathbf{u}_i,$$

   where $\mathbf{u}_i$ is a normally distributed random variable with mean 0 and standard deviation σ chosen such that the adjusted coefficient of determination is approximately $R^2 = 0.5$.
3. We drew 5,000 repeated samples from the target population and calculated the mean square error of each estimator of the total t_{y^*}.
4. Steps 1-3 were repeated by fitting a lasso regression model with main effects and 2-way interactions and a regression tree model using the algorithm detailed in Section 2.5.

Table 4.2 displays the ratio the design MSE of each estimator to that of the HT under the three different models generating the survey variable of interest for a sample size of $n = 1{,}000$. As expected, the estimator based on the correctly specified working model is the most efficient. In the case where the true generating model contains only main effects, assuming a working model with higher order interactions results in a slight loss in efficiency. If two-way or higher order interactions are present, the regression tree and lasso-based estimators fitted with two-way interactions are more efficient than the model-assisted estimators based on working models with only main effects. When the generating model is a regression tree, the regression tree estimator yields modest efficiency gains over the 2-way lasso-based estimators. This can be explained by the fact that the regression tree model groups the categories of an auxiliary variable based on their relationship to the variable of interest and, therefore, reduces the model size. In all cases, significant efficiency gains, relative to the design-based HT estimator, are achieved.

Table 4.2
Ratio of MSE for each estimator to MSE of HT under different models generating survey variable of interest

	LASSO (1-way)	LASSO (2-way)	Regression Tree
GREG	0.749	0.855	0.878
FSTEP	0.749	0.855	0.876
TREE	0.803	0.821	0.778
LASSO (1-way)	0.747	0.850	0.871
CLASSO (1-way)	0.747	0.851	0.873
LASSO (2-way)	0.763	0.761	0.826
CLASSO (2-way)	0.763	0.765	0.833
ALASSO	0.750	0.849	0.872
CALASSO	0.750	0.851	0.873

4.2 Performance under other scenarios

We also examined the performance of the lasso-based and regression tree estimators under scenarios where there are no main effects, only 2-way interactions. We generated a fourth survey variable of interest

using the lasso regression model with main effects and 2-way interactions as described in the procedure above. However, in step 2, we set all coefficients estimates corresponding to main effects equal to 0.

The first column of Table 4.3 (called no multicollinearity) shows the ratio the design MSE of the estimators to that of the HT estimator, where the survey variable is generated from a model with no main effects for sample sizes of $n = 1{,}000$. Under this scenario, the lasso estimators with 2-way interactions and the regression tree estimator are significantly more efficient than model-assisted estimators based on main effects only models. Relative to the commonly used GREG estimator, the efficiency gains for the lasso estimators with 2-way interactions and the regression tree estimator are significantly greater when there are no main effects. This is evident by comparing LASSO 2-way column in Table 4.2 to first column in Table 4.3. The relative MSE is very similar for the 2-way lasso and regression tree estimators but closer to 1 for GREG and 1-way lasso estimators.

Table 4.3
Ratio of MSE for each estimator to MSE of HT under generating model with no main effects and in the absence/presence of multicollinearity

	No Multicollinearity	Duplicated Variable	Collapsed Categories
GREG	0.935	-	-
TREE	0.824	0.850	0.842
LASSO (1-way)	0.930	0.945	0.942
CLASSO (1-way)	0.936	0.953	0.951
LASSO (2-way)	0.783	0.795	0.773
CLASSO (2-way)	0.795	0.809	0.781

For administrative data with many variables, it is not uncommon for some variables to be colinear or nearly colinear. For example, information on both the total number of employees and the number of full-time equivalent employees is often available. The GREG estimator, and by extension the FSTEP estimator and adaptive lasso estimators, fail in the presence of collinearity as the design matrix is singular. We investigated the performance for regression tree and lasso estimators in the presence of multicollinearity. We considered two types of multicollinearity:

- Duplicate of existing categorical variable. We created three new indicator variables corresponding to employment size.
- Collapsed categories of existing auxiliary variable: We created a new indicator variable corresponding to the three highest categories of revenue.

The MSE, relative to the HT estimator, for $n = 1{,}000$ is shown in columns 2 and 3 of Table 4.3. These results are very similar to those in the first column of Table 4.3 without the presence of multicollinearity. The regression tree and lasso estimators provide an automatic way of removing colinear auxiliary variables without impacting the potential efficiency gains. It should be noted that other methods, such as principal component analysis, can be used to eliminate collinearity but require some expertise.

4.3 Performance of variance estimators in terms of relative bias

Variance estimators based on (2.8) were constructed for each estimator. Table 4.4 displays the percentage relative bias of each estimator for the total amount of trade credit requested. For comparison purposes, the theoretically unbiased variance estimator of the HT estimator is included in this table. This variance estimator is equivalent to the expression provided in (2.8) where $e_i = y_i - \bar{y}_s$. The variance estimators for the model-assisted survey regression estimators have substantial negative bias which increases as the number of auxiliary variables, p, increases. The magnitude of negative bias is largest for the lasso-based estimators fitted using 2-way interactions. For small sample sizes, the negative bias is smallest for the regression tree estimator. As well, for small sample sizes, there is a substantial difference in bias between the GREG and FSTEP estimators. Performing variable selection prior to calculating the standard GREG calibration estimator appears to reduce the bias of the variance estimator in this case. The bias reduces for all model-assisted survey regression estimators as the sample size increases.

Table 4.4
Percent relative bias of variance estimators

	20 categories			28 categories		
	$n = 200$	$n = 500$	$n = 1{,}000$	$n = 200$	$n = 500$	$n = 1{,}000$
GREG	-12.44	-4.16	-1.60	-22.23	-10.86	-6.99
FSTEP	-7.05	-3.60	-1.62	-14.07	-7.71	-6.73
TREE	-5.79	-5.53	-2.81	-8.45	-12.93	-10.83
LASSO (1-way)	-7.79	-2.96	-1.14	-12.42	-9.49	-6.44
CLASSO (1-way)	-10.08	-3.74	-1.61	-16.01	-9.84	-6.52
LASSO (2-way)	-11.94	-11.57	-7.62	-16.12	-15.14	-13.08
CLASSO (2-way)	-19.99	-15.09	-9.06	-25.87	-19.04	-15.14
ALASSO	-8.69	-3.61	-1.41	-14.52	-9.43	-6.38
CALASSO	-9.40	-3.78	-1.48	-15.80	-9.64	-6.46
HT	5.19	5.72	5.82	4.90	-0.11	1.66

Given the bias of the variance estimators seen here, particularly for small sample sizes, a possible concern is the quality of the first-order Taylor expansion approximation. For a large number of categorical auxiliary variables, the remainder term in the Taylor expansion may no longer be negligible for small sample sizes. An alternative variance estimator for the lasso estimators was considered by McConville et al. (2017) but yielded only slight improvements in terms of bias reduction. An additional concern is properly accounting for the inherently data driven procedure used to estimate the regression tree and lasso models. The regression tree model has splits while the lasso models have a penalty parameter both depending on the sample.

4.4 Properties of the survey weights

Regression weights are directly available for the GREG, FSTEP, regression tree, lasso calibration (1-way and 2-way) and adaptive lasso calibration estimators. We investigated the properties of the weights for these estimators in our simulations.

Large variation in the values of weights is undesirable as they allow some units to be much more influential than others. Positive weights are preferred by national statistical organizations as a negative weight no longer holds the interpretation of the number of population units represented by the sampled unit.

First, we computed the average, over repeated samples, of the empirical within-sample variance of the weights:

$$\overline{\text{var}}(\mathbf{w}) = \frac{1}{R}\sum_{r=1}^{R}\frac{1}{n-1}\sum_{j\in s^{(r)}}\left(w_j^{(r)} - \bar{w}^{(r)}\right)^2,$$

where $s^{(r)}$ is the rth simulated sample, $\bar{w}^{(r)} = \frac{1}{n}\sum_{j\in s^{(r)}} w_j^{(r)}$ and $w_j^{(r)}$ is the weight of the jth unit in the rth simulated sample.We also computed the average coefficient of variation (CV) of the weights:

$$\overline{\text{CV}}(\mathbf{w}) = \frac{1}{R}\sum_{r=1}^{R}\frac{\sqrt{\frac{1}{n-1}\sum_{j\in s^{(r)}}\left(w_j^{(r)} - \bar{w}^{(r)}\right)^2}}{\bar{w}^{(r)}}$$

Table 4.5 displays the average variance and average CV for the weights across samples when revenue was included as an auxiliary variable. The weights for the GREG estimator and, to a lesser extent the FSTEP estimator, are much more variable than the weights for the regression tree and lasso-based estimators, particularly for small sample sizes. The variability of the weights for the three lasso-based approaches is very similar and is always slightly lower than the variability of the weights for the regression tree estimator.

Table 4.5
Average variance (CV) for weights across samples

	$n = 200$	$n = 500$	$n = 1{,}000$
GREG	728.18 (0.59)	77.14 (0.48)	16.41 (0.44)
FSTEP	462.81 (0.47)	67.45 (0.45)	15.90 (0.44)
TREE	374.43 (0.42)	59.35 (0.42)	14.70 (0.42)
CLASSO (1-way)	354.57 (0.41)	56.21 (0.41)	14.03 (0.41)
CLASSO (2-way)	361.83 (0.42)	56.60 (0.41)	14.06 (0.41)
CALASSO	354.29 (0.41)	56.28 (0.41)	14.03 (0.41)

We also computed the proportion of simulated samples where the regression weights contained negative values. As mentioned in Section 2.5, by construction, the weights for the regression tree estimator are guaranteed to be strictly positive. When the sample size was 200, the GREG estimator calibrated to 20 marginal categories yielded negative weights for approximately 3% of the repeated samples. There were no negative weights when the sample size was 500 or 1,000. For the GREG estimator calibrated to 28 marginal categories, approximately 27% of the repeated samples of size 200 contained negative weights and less than 0.5% of the repeated samples of size 500 contained negative weights. The GREG weights are unstable when the sample size is small, especially if the GREG estimator is calibrated

to auxiliary variables with many categories. Using forward stepwise variable selection with the GREG estimator resulted in a substantial decrease in the number of simulated samples with negative weights for small sample sizes. The FSTEP estimator applied to the 28 marginal categories yielded negative weights in approximately 0.5% of the repeated samples of size 200. There were no negative weights observed for the lasso calibration estimator with only main effects or adaptive lasso calibration estimator. Using the lasso calibration estimator with 2-way interactions resulted in negative weights in less than 0.05% of the simulated samples.

4.5 Estimation based on a single set of weights

A major drawback in the implementation of the regression tree and the calibrated lasso-based approaches is that the estimation procedures yield variable-specific weights. We conducted additional simulations in which a single set of variable-specific weights was applied to other related survey variables of interest. In the context our business survey data, we considered four survey variables of interest, the amount of trade credit requested as well as the amount requested for three additional types of financing: line of credit, business credit card and leasing financing. We examined the impact on bias and loss of efficiency in using a single set of weights, determined by a primary variable of interest, to estimate the total amount requested for the remaining three survey variables of interest. Specifically, we calculated the percentage absolute relative design bias for the estimators of the total amount requested and the variance estimators. We also calculated the ratio of the MSE for the regression tree and three calibrated lasso-based approaches using the set of weights corresponding to a primary variable of interest to the MSE for the estimators using variable-specific weights. For brevity, we considered only settings with 28 marginal categories.

The percentage absolute relative design bias was less than 2 percent for all of the estimators for all scenarios. For all estimators and primary variable of interest, the bias decreases as the sample size increases.

Unlike the bias of the variance estimators based on variable-specific weights, the bias of the variance estimators based on a single set of weights for a primary variable of interest does not necessarily decrease as the sample size increases. As well, the bias is not strictly in one direction and may be positive or negative. For the regression tree and calibrated lasso-based approaches, the bias of the variance estimators is substantially larger for the primary variable of interest used to calculate the single set of weights than for the other study variables. The data driven nature of these estimators means that the estimated variance for the primary variable of interest is underestimated, as shown in Table 4.4.

Table 4.6 displays the ratio of the design MSE of each estimator with weights determined by a primary variable of interest to that of the estimator with variable-specific weights, calculated separately for each of the four study variables for n equal to 200 and 500. Using a single set of weights determined by a primary variable of interest results in a similar or slightly higher MSE than using variable-specific weights. Here,

the loss in efficiency is modest, less than 8% in all settings considered. Similar results were obtained for the case $n = 1{,}000$. There is no clear pattern in terms of loss of efficiency and sample size.

Table 4.6
Ratio of MSE for each estimator with weights determined by primary variable of interest to MSE for estimator with variable-specific weights

		Trade Credit		Line of Credit		Business Credit Card		Lease Financing	
	n	**200**	**500**	**200**	**500**	**200**	**500**	**200**	**500**
Primary variable: Trade Credit	TREE	-	-	1.01	0.97	0.99	1.00	0.99	1.00
	CLASSO (1-way)	-	-	0.99	0.99	1.01	0.99	1.00	1.01
	CLASSO (2-way)	-	-	0.93	0.94	0.92	0.98	0.92	0.97
	CALASSO	-	-	0.97	0.99	1.01	0.99	0.96	1.00
Primary variable: Line of Credit	TREE	1.06	0.97	-	-	0.98	1.00	0.98	0.97
	CLASSO (1-way)	0.96	0.98	-	-	0.99	1.01	0.99	0.99
	CLASSO (2-way)	0.95	0.96	-	-	0.92	0.98	0.93	0.96
	CALASSO	0.97	0.98	-	-	0.99	1.00	0.96	0.98
Primary variable: Business Credit Card	TREE	1.06	1.01	1.06	0.97	-	-	0.99	1.02
	CLASSO (1-way)	0.99	1.02	0.98	0.97	-	-	0.99	1.02
	CLASSO (2-way)	0.98	1.00	0.95	0.93	-	-	0.92	0.99
	CALASSO	1.00	1.02	0.97	0.97	-	-	1.00	1.01
Primary variable: Lease Financing	TREE	1.07	1.03	1.06	1.05	0.99	1.02	-	-
	CLASSO (1-way)	0.99	1.05	0.98	1.04	0.99	1.02	-	-
	CLASSO (2-way)	0.97	1.02	0.96	1.01	0.92	0.99	-	-
	CALASSO	1.00	1.05	0.98	1.05	1.00	1.01	-	-

5. Estimation under non-probability sampling

In this section, we study the effect of selection bias on the survey regression estimators under non-probability sampling. For this purpose, we studied two types of selection bias possibly present in non-probability samples. In particular, we considered a scenario in which the probability of selection depends only on the auxiliary data available for all units in the population, and a scenario in which the probability of selection depends on the survey variable of interest. In both scenarios, we evaluated the absolute relative bias (ARB), $\left|\hat{t}_y - t_y\right| \big/ t_y$, for each estimator of the total. Following Chen, Valliant and Elliott (2018), we treat the non-probability sample as a simple random sample and set the design weights equal to $d_i = N/n$ for the estimation of total t_y as the selection process for non-probability samples is unknown in practice.

5.1 Selection probabilities depend on auxiliary data

We drew repeated samples using the same stratified SRS design as in Section 4. Table 5.1 displays the ARB of each estimator of the total amount of trade credit requested assuming $d_i = N/n$, when the sample is in fact selected using disproportionate stratified random sampling.

As expected, the wholly designed-based HT estimator has the largest bias, and this bias does not decrease as the sample size increases. The ARB of model-assisted estimators decreases as the sample size

n increases. The GREG estimator has the smallest bias, particularly for small sample sizes. Furthermore, the GREG estimator is approximately unbiased if revenue is included as one of the auxiliary variables for calibration. However, if stepwise variable selection is used, the GREG estimator is no longer unbiased for small sample sizes. On the other hand, if revenue is not included as a calibration variable, the GREG estimator is slightly biased. The lasso-based and, to a smaller extent, the regression tree estimators suffer from small sample bias for $n = 200$ when revenue is correctly included as an auxiliary variable. This is most apparent for the standard lasso estimators that do not include calibration to known population totals. For n equal to 500 or 1,000, including revenue as an auxiliary variable, substantially decreases the bias for the regression tree and calibrated lasso estimators but only slightly decreases the bias for the lasso estimators without calibration. This indicates that the additional calibration step is important for diminishing the effect of selection bias, especially if the sample size is small.

Table 5.1
Percent ARB of each estimator under stratified sampling with revenue and without revenue included as an auxiliary variable

	Revenue			Without Revenue		
	$n = 200$	$n = 500$	$n = 1{,}000$	$n = 200$	$n = 500$	$n = 1{,}000$
GREG	0.31	0.06	0.06	4.84	5.12	4.71
FSTEP	2.67	0.44	0.06	9.20	5.18	4.92
TREE	4.15	1.04	0.50	17.40	10.20	8.94
LASSO (1-way)	17.42	5.10	2.32	16.32	8.88	6.49
CLASSO (1-way)	7.99	0.83	0.20	9.04	5.22	4.59
LASSO (2-way)	25.36	14.28	8.40	26.31	15.16	9.89
CLASSO (2-way)	10.72	1.44	1.02	14.19	5.56	3.84
ALASSO	14.95	5.63	3.00	14.35	8.64	6.51
CALASSO	9.63	2.54	1.25	9.27	5.77	4.92
HT	49.45	48.84	48.81	49.08	49.29	48.60

These results indicate that when the selection probability depends on a known auxiliary variable, including it in the working model for the GREG estimator effectively diminishes the effect of selection bias. This was not the case for the model-assisted estimators that involved variable selection. Performing variable selection may increase bias as auxiliary variables that are predictive in terms of selection probability may not be selected and properly accounted for. The lasso estimators can be constructed such that user-specified variables are always included in the working regression model. These user-specified variables can be added to x_i^* in equation (2.5) to force calibration to corresponding population totals. Unfortunately, the underlying selection mechanism is unknown in practice and, therefore, correctly identifying variables which impact selection probability is challenging.

5.2 Selection probabilities depend on the study variable

Next, we drew repeated samples using Poisson sampling where the sampling probabilities depends on the survey variable of interest. We assume the Poisson sampling probabilities are given by:

$$\text{logit}\left(p_i\right) = \beta_0 + \beta_1 y_i$$

where y_i is the amount of trade credit requested in millions of dollars, $\beta_1 = 0.5$ and $\beta_0 = -3.80, -2.85, -2.10.$ The intercept values, β_0, were chosen such that we obtained sample sizes of approximately 200, 500 and 1,000 units, averaged over the simulated samples. Under this sampling design, units with larger amounts requested for trade credit have a higher probability of being sampled and, therefore, are over-represented. Table 5.2 displays the ARB of each estimator of the total amount of trade credit requested assuming $d_i = N/n$, when the sample is selected using the above informative Poisson sampling. Here, all the estimators are heavily biased because the population model does not hold due to informative sampling. The magnitude of the bias is very similar across estimators and does not substantially decrease as the sample size increases. The inclusion or exclusion of revenue as an auxiliary variable does not impact the bias.

Table 5.2
Percent ARB of each estimator under Poisson sampling with revenue and without revenue included as an auxiliary variable

	Revenue			Without Revenue		
	β_0 = -3.8	β_0 = -2.85	β_0 = -2.1	β_0 = -3.8	β_0 = -2.85	β_0 = -2.1
GREG	23.53	22.27	20.45	24.74	22.91	21.21
FSTEP	24.54	22.55	20.58	25.16	23.24	21.15
TREE	24.07	22.73	20.15	24.93	22.47	20.55
LASSO (1-way)	24.29	22.73	20.65	25.45	23.29	21.38
CLASSO (1-way)	23.02	22.30	20.47	24.74	22.99	21.23
LASSO (2-way)	23.15	22.06	20.17	24.66	22.73	20.62
CLASSO (2-way)	20.11	20.18	19.01	22.62	21.63	19.98
ALASSO	24.44	22.72	20.66	25.50	23.21	21.36
CALASSO	23.91	22.46	20.53	25.10	23.01	21.25
HT	29.12	27.95	25.57	29.36	27.53	25.45

6. Conclusions

We have evaluated the performance of several model-assisted survey regression estimators, in the context of both probability and non-probability sampling, through a simulation study. First, we discuss the overall conclusions from our simulation study using probability samples with a stratified SRS design. In the context of our business survey data with all categorical auxiliary variables, the regression tree estimator and the lasso (2-way) estimator with two factor interaction effects are the only model-assisted estimators that provide any efficiency gains, relative to the HT estimator, when the sample size is small and the number of categories of auxiliary variables used is large. As well, the variance estimator for the regression tree estimator is the least biased in this scenario. As the sample size increases, the difference in efficiency between the model-assisted survey regression estimators becomes negligible and all are slightly more efficient than the HT estimator. In general, the potential gains in efficiency for model-assisted estimators over the HT estimator depend on the predictive power of the model. In our simulation

population, the strength of the relationship between the study variable and the available categorical auxiliary variables is somewhat weak as judged by the adjusted coefficient of determination R^2 around 0.20. We therefore generated study variables leading to larger R^2 values around 0.50 by making the model error variance smaller. As expected, model-assisted estimators led to significant efficiency gains over the HT estimator in all cases, as reported in Table 4.2 which shows that the regression tree estimator and the lasso estimator with interaction effects yield improved efficiency over the commonly used GREG estimator if two-factor interactions are present. Moreover, the regression weights for the tree estimator and the calibration weights for the lasso calibration estimators are much less variable, particularly for small sample sizes, than the weights for the GREG. We also examined the performance of the lasso-based and regression trees estimators under a scenario with no main effects and only two-factor interactions are present and another scenario where multi-collinearity among the auxiliary variables is present. In the latter scenario, GREG is not applicable, and we show that the regression tree and lasso estimators provide an automatic way of removing colinear auxiliary variables without impacting the potential efficiency gains. Overall, we recommend using either lasso (2-way) or regression tree estimators in terms of efficiency when two factor interactions are likely to be present among the categorical auxiliary variables. Even in the case of models with only main effects, both methods perform well relative to GREG in terms of MSE because the lasso (2-way) estimator automatically shrinks regression coefficients associated with the interactions to zero while the regression tree estimator does not require specification of the mean function. In other contexts where there is evidence of complex non-linear and non-additive relationships between the survey variable of interest and auxiliary variables, the use of other tree-based machine learning methods, such as xgboost and random forests, should be studied.

In Section 4.3, we studied the performance of variance estimators in terms of relative bias and showed that all the variance estimators exhibit significant underestimation for sample size $n = 200$ and 28 x-categories. Relative bias of the regression tree variance estimator did not decrease as the sample size increased, unlike in the other cases, and it could be due to overfitting. In the context of random forests method, Dagdoug, Goga and Haziza (2021) examined a procedure based on cross-validation which led to small relative biases and good coverage rates. It would be worthwhile to study a similar procedure for variance estimation of the regression tree estimator.

A major drawback of the regression tree and lasso-based approaches is that the estimation procedures do not yield a set of generic weights that can be applied to all study variables, y. A possible alternative approach is to derive regression weights based on a primary variable of interest and apply that set of weights to related study variables. In the survey context considered here, using a single set of weights for a group of related variables resulted in little loss of efficiency, relative to the use of variable-specific weights. As well, the bias of the estimators remained negligible. Under this approach, the desirable properties of the regression weights, low variability and, in the case of the regression tree estimator, strictly positive weights are maintained. However, the asymptotic properties of the lasso and regression

tree survey estimators have not been derived for a single set of weights, applied to multiple study variables.

We also considered the use of model-assisted survey regression estimators for data from mis-specified probability sampling, treated as a non-probability sample. When the probability of selection depends on an observed auxiliary variable, the bias of the model-assisted estimators decreases as the sample size increases. Including the appropriate auxiliary variable in the working model for the GREG estimator effectively removes the selection bias. Achieving this in practice is difficult as the selection process is unknown. Performing variable selection can increase the bias for model-assisted survey regression estimators as the auxiliary variables related to the selection probability may not be included in the regression model. In fact, in our simulations, correctly including revenue as a potential auxiliary variable did not necessarily decrease the bias of the lasso estimators.

When the probability of selection depends on the survey variable of interest, all the estimators are heavily biased. The magnitude of the bias is similar across estimators and does not greatly decrease as the sample size increases. In our simulation population, the auxiliary variables are not highly predictive for the survey variables of interest. Examining the impact of the strength of the relationship between the auxiliary variables and the variable of interest when informative selection is present warrants more investigation.

Sample selection bias may not be reduced by using a non-probability sample alone, as demonstrated in our simulation study. Methods based on integrating a non-probability sample observing the study variables and associated auxiliary variables with a probability sample observing only the same auxiliary variables have the potential of reducing selection bias through modeling the participation probabilities (Chen, Li and Wu, 2020). Dual frame screening methods are also available when the study variable is observed in both samples and the units in the probability sample belonging to the non-probability sample can be identified without linkage errors without the need to model the participation probabilities (Kim and Tam, 2020; Rao, 2021 and Beaumont, 2020). However, the dual frame method is effective only when the sampling fraction for the non-probability sample is large. We are studying the above methods in the context of business surveys, for example integrating survey data with incomplete administrative data treated as a non-probability sample.

Acknowledgements

We thank Dr. Wesley Yung for initiating this work and for his constructive comments and suggestions. We also thank the reviewers, the editor and the associate editor for constructive comments and suggestions.

References

Beaumont, J.-F. (2020). Are probability surveys bound to disappear for the production of official statistics? *Survey Methodology*, 46, 1, 1-28. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2020001/article/00001-eng.pdf.

Breidt, F.J., and Opsomer, J.D. (2017). Model-assisted survey estimation with modern prediction techniques. *Statistical Science*, 32(2), 190-205.

Buskirk, T.D., Kirchner, A., Eck, A. and Signorino, C.S. (2018). An introduction to machine learning methods for survey researchers. *Survey Practice*, 11(1), 1-10.

Cassel, C.M., Särndal, C.-E. and Wretman, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite population. *Biometrika*, 63(3), 615-620.

Chen, Y., Li, P. and Wu, C. (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association*, 115(523), 2011-2021.

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.

Chen, J.K.T., Valliant, R.L. and Elliott, M.R. (2019). Calibrating non-probability surveys to estimated control totals using LASSO, with an application to political polling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 68(3), 657-681.

Dagdoug, M., Goga, C. and Haziza, D. (2021). Model-assisted estimation through random forests infinite population sampling. *Journal of the American Statistical Association* (to appear).

Friedman, J., Hastie, T., Simon, N., Qian, J. and Tibshirani, R. (2017). *glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models.* R package version 2.0-13.

Kern, C., Klausch, T. and Kreuter, F. (2019). Tree-based machine learning methods for survey research. *Survey Research Methods*, 13(1), 73-93.

Kern, C., Li, Y. and Wang, L. (2020). Boosted kernel weighting-using statistical learning to improve inference from nonprobability samples. *Journal of Survey Statistics and Methodology*. https://doi.org/10.1093/jssam/smaa028.

Kim, J.K., and Tam, S.M. (2020). Data integration combing big data and survey sample data for finite population inference. *International Statistical Review*, 89(2), 382-401.

McConville, K.S. (2011). *Department of Statistics Improved Estimation for Complex Surveys Using Modern Regression Techniques*, unpublished Ph.D. thesis, Colorado State University.

McConville, K.S., and Toth, D. (2019). Automated selection of post-strata using a model-assisted regression tree estimator. *Scandinavian Journal of Statistics*, 46(2), 389-413.

McConville, K.S., Breidt, F.J., Lee, T.C.M. and Moisen, G.G. (2017). Model-assisted survey regression estimation with the LASSO. *Journal of Survey Statistics and Methodology*, 5(2), 131-158.

McConville, K.S., Tang, B., Zhu, G., Li, S., Cheung, S. and Toth, D. (2018). *mase: Model-Assisted Survey Estimators*. R package version 0.1.1.

Rafei, A., Flannagan, C.A. and Elliott, M.R. (2020). Big data for finite population inference: Applying quasi-random approaches to naturalistic driving data using Bayesian additive regression trees. *Journal of Survey Statistics and Methodology*, 8(1), 148-180.

Rao, J.N.K. (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhyā B*, 83(1), 242-272 (published online April 2020).

Ripley, B., Venables, B., Bates, D.M., Hornik, K., Gebhardt, A. and Firth, D. (2017). *MASS: Modern Applied Statistics with S.* R package version 7. 3-47.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B*, 58(1), 267-288.

Zou, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, 101(476), 1418-1429.

Survey Methodology, June 2022
Vol. 48, No. 1, pp. 73-93
Statistics Canada, Catalogue No. 12-001-X

Bayesian inference for a variance component model using pairwise composite likelihood with survey data

Mary E. Thompson, Joseph Sedransk, Junhan Fang and Grace Y. Yi[1]

Abstract

We consider an intercept only linear random effects model for analysis of data from a two stage cluster sampling design. At the first stage a simple random sample of clusters is drawn, and at the second stage a simple random sample of elementary units is taken within each selected cluster. The response variable is assumed to consist of a cluster-level random effect plus an independent error term with known variance. The objects of inference are the mean of the outcome variable and the random effect variance. With a more complex two stage sampling design, the use of an approach based on an estimated pairwise composite likelihood function has appealing properties. Our purpose is to use our simpler context to compare the results of likelihood inference with inference based on a pairwise composite likelihood function that is treated as an approximate likelihood, in particular treated as the likelihood component in Bayesian inference. In order to provide credible intervals having frequentist coverage close to nominal values, the pairwise composite likelihood function and corresponding posterior density need modification, such as a curvature adjustment. Through simulation studies, we investigate the performance of an adjustment proposed in the literature, and find that it works well for the mean but provides credible intervals for the random effect variance that suffer from under-coverage. We propose possible future directions including extensions to the case of a complex design.

Key Words: Cluster sample analysis; Composite likelihood; Curvature adjustment; Random effects model.

1. Introduction

Multi-stage survey designs are used in many population-based surveys. Increasingly, multilevel models have been used to make inferences when data are obtained from a multi-stage survey.

Desiring to improve such inferences Rao, Verret and Hidiroglou (2013) (RVH) proposed using a weighted log pairwise composite likelihood approach. There is an extensive literature on composite likelihoods: see review papers by Varin (2008), Varin, Reid and Firth (2011) and Yi (2017), and many applications. In their Section 4 RVH describe a unified approach applicable to both linear and generalized linear models. Important aspects of their work include (a) obtaining design-consistent point estimates of mean and regression parameters and variance components, and (b) using only first-order inclusion probabilities and second-order probabilities within clusters. In particular, RVH work in (a) is important because of design inconsistency when the number of clusters (first-stage units) grows while the cluster sample sizes remain small (Pfeffermann, Skinner, Holmes, Goldstein and Rasbash, 1998). Unlike the pseudo-likelihood approach in common use (Rabe-Hesketh and Skrondal, 2006) their method ensures that (a) holds for outcomes from generalized linear models. The research in RVH has been extended by Yi, Rao and Li (2016) (YRL), who provide a more general framework, additional theory and extensive simulations.

1. Mary E. Thompson, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada. E-mail: methompson@uwaterloo.ca; Joseph Sedransk, Joint Program in Survey Methodology, University of Maryland, College Park, MD 20742, U.S.A.; Junhan Fang, School of Medicine, Yale University, New Haven, CT 06520, U.S.A.; Grace Y. Yi, Department of Statistical and Actuarial Sciences, Department of Computer Science, Western University, London, Ontario, N6A 5B7, Canada.

Two related developments have led to our research. First, there is increasing interest in the use of Bayesian methods for inferences from survey data. Section 5 has a general reference together with an introduction to papers describing extensive use of Bayesian methods at the National Agricultural Statistical Service of the US Department of Agriculture. Second, there is (Bayesian) literature demonstrating the possibility of overstated precision by using *unadjusted* composite likelihoods, e.g., Ribatet, Cooley and Davison (2012) (RCD) and Stoehr and Friel (2018).

Our approach is to start with a posterior distribution taken proportional to the product of a composite likelihood and prior distribution. Comparing this approximate posterior distribution with one using the full likelihood, we show that inferences based on the approximate posterior exhibit overstated precision. Making adjustments to the posterior distribution based on the composite likelihood as in RCD, we then use simulations to compare the three ways of formulating a posterior distribution, i.e., those based on the full, composite and adjusted composite likelihoods. This is done by visual evaluation of the graphs of the posterior densities and coverages (over repeated simulations) of 95% credible intervals for the model parameters.

The methodology is described in Section 2.3. The adjustments to the approximate posterior distribution based on a composite likelihood are derived from a transformation of the logarithm of the composite likelihood at its mode, designed so that the negative of the inverse of the curvature matrix of the approximate posterior density at its mode will match the corresponding posterior variance-covariance matrix of the parameters. This is similar to the property in frequentist inference that the inverse of the observed Fisher information matrix (negative of the Hessian of the log likelihood at its mode) estimates the variance-covariance matrix of the maximum likelihood estimates.

To focus on the main issue we use a "noninformative" prior distribution for the parameters of our model, described below. Then the corresponding posterior density is close to the normalized likelihood, and advances shown in a Bayesian context would also be seen in a frequentist model-based approach.

To simplify the initial investigation a standard linear random effects (intercept only) superpopulation model is assumed. Consider a survey population drawn from that superpopulation and composed of a large number N of clusters, each of a common size, say, m. Let Y_{ij} denote the continuous response variable for elementary unit j in cluster i with $i = 1, \ldots, N$ and $j = 1, \ldots, m.$ Then we write

$$Y_{ij} = \theta + u_i + e_{ij} \tag{1.1}$$

where $u_i \sim N\left(0, \sigma_u^2\right), e_{ij} \sim N\left(0, \sigma_e^2\right),$ all u_i and e_{ij} are independent, and θ, σ_u, and σ_e are parameters.

We also begin by assuming that the survey sampling design is a simple random sample of n clusters, where n is a positive integer. This has the advantage that the model (1.1) holds not only for the superpopulation and the finite population but also (replacing N by n) for the sample itself, arising from generation of the population followed by the selection of the sample using the sampling design. It ensures that the likelihood function to be used in Bayesian inference is well defined. As well, it can be shown that

Bayesian inference from the sample for the parameters of model (1.1) would be interpretable also in terms of the frequentist theory for analytic uses of survey data (Skinner, Holt and Smith, 1989).

Our work is valuable because we show the perils of using an *unadjusted* pairwise composite likelihood to form an approximate posterior distribution for inference even in this very simple and straightforward case. Extensions to unequal probability sampling designs are discussed in Section 4.

The proposed adjustment leads to excellent frequentist properties for inference about the mean θ. The posterior mean of θ has low bias in the frequency sense, and the frequentist coverage of credible intervals aligns with the nominal levels. For σ_u it provides a significant improvement over using the unadjusted composite likelihood. However, the coverage falls short of the nominal level, leading to the need for additional research about how to modify the adjustment.

The rest of the paper is structured as follows. Section 2 provides the definitions of the full, composite and adjusted composite likelihoods and the prior distributions. There follows a description of the curvature adjustment and the reasons for its use. The simulation studies are described in Section 3 including the model, prior distributions, sample sizes and their settings, number of replications, etc. This section also describes how the results are displayed together with a summary of our findings. Extensions to unequal probability sampling cases are discussed in Section 4. Conclusions are in Section 5.

2. Full likelihood, pairwise likelihood and Bayesian implementation

2.1 Model and formulae

As in Section 1, let Y_{ij} denote the response variable for second-stage unit j in first-stage unit i for $i = 1, \ldots, n,$ and $j = 1, \ldots, m.$ We use lower case letter y_{ij} to represent realized values of Y_{ij}. Let $\mathbf{y}(n) = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ denote the sample data with $\mathbf{y}_i = (y_{i1}, \ldots, y_{im})^{\mathrm{T}}$ for $i = 1, \ldots, n,$ where T denotes transpose.

In a more general random effects model, we might assume that, conditional on random effects u_i for $i = 1, \ldots, n,$ the Y_{ij} are independently distributed as

$$Y_{ij} \sim f_{y|u}\left(y_{ij} \,\middle|\, u_i; \boldsymbol{\theta}_y\right) \quad \text{for} \quad j = 1, \ldots, m, \tag{2.1}$$

where $f_{y|u}$ is a known density function and $\boldsymbol{\theta}_y$ is the associated parameter vector. Next, we model random effects by assuming that the u_i are independent and identically distributed as

$$u_i \sim f_u\left(u_i \,\middle|\, \boldsymbol{\theta}_u\right) \quad \text{for} \quad i = 1, \ldots, n, \tag{2.2}$$

where f_u is a given density function indexed by the parameter vector $\boldsymbol{\theta}_u$.

Let $\boldsymbol{\eta} = \left(\boldsymbol{\theta}_y^{\mathrm{T}}, \boldsymbol{\theta}_u^{\mathrm{T}}\right)^{\mathrm{T}}$ be the vector of model parameters which is of interest. In the frequentist framework, the maximum likelihood method is commonly used to conduct inference about $\boldsymbol{\eta}$ by maximizing the likelihood function

$$L(\boldsymbol{\eta}) = \prod_{i=1}^{n} f(\mathbf{y}_i; \boldsymbol{\eta}),$$

where

$$f(\mathbf{y}_i; \boldsymbol{\eta}) = \int \left\{ \prod_{j=1}^{m_i} f_{y|u}\left(y_{ij} \,|\, u_i; \boldsymbol{\theta}_y\right) \right\} f_u\left(u_i \,|\, \boldsymbol{\theta}_u\right) du_i. \tag{2.3}$$

An alternative to the likelihood method is the composite likelihood approach (Lindsay, 1988). In particular, the pairwise likelihood method has often been employed. Let $L_{ij}(\boldsymbol{\eta}) = f\left(y_{ij}; \boldsymbol{\eta}\right)$ be the density of Y_{ij}, determined by

$$f\left(y_{ij}; \boldsymbol{\eta}\right) = \int f_{y|u}\left(y_{ij} \,|\, u_i; \boldsymbol{\theta}_y\right) f_u\left(u_i \,|\, \boldsymbol{\theta}_u\right) du_i.$$

For $j \neq k$, let $L_{ijk}(\boldsymbol{\eta}) = f\left(y_{ij}, y_{ik}; \boldsymbol{\eta}\right)$ be the joint density for paired responses $\left(Y_{ij}, Y_{ik}\right)$, determined by

$$f\left(y_{ij}, y_{ik}; \boldsymbol{\eta}\right) = \int f_{y|u}\left(y_{ij} \,|\, u_i; \boldsymbol{\theta}_y\right) f_{y|u}\left(y_{ik} \,|\, u_i; \boldsymbol{\theta}_y\right) f_u\left(u_i \,|\, \boldsymbol{\theta}_u\right) du_i.$$

Then a marginal pairwise likelihood function can be formulated as

$$C(\boldsymbol{\eta}) = \prod_{i=1}^{n} \prod_{j<k} L_{ijk}^{d_{jk}}(\boldsymbol{\eta}) \times L_{ij}^{d_j}(\boldsymbol{\eta}) \times L_{ik}^{d_k}(\boldsymbol{\eta}),$$

where d_{jk}, d_j, and d_k are weights that can be user-specified to enhance efficiency or to facilitate some specific features of the formulation. Discussion on choosing weights can be found in Cox and Reid (2004), Lindsay, Yi and Sun (2011), Varin, Reid and Firth (2011), and Yi (2017). To confine our attention to the use of marginal pairwise likelihoods, in line with the approach of RVH, here we consider the case with $d_j = d_k = 0$ and $d_{jk} = 1$.

Returning to the special case of model (1.1), suppose that σ_e^2 is known, and take $\boldsymbol{\eta}$ to consist of $\boldsymbol{\theta}_y = \theta$ and $\boldsymbol{\theta}_u = \sigma_u^2$. In a Bayesian approach it is necessary to choose a prior distribution for $\boldsymbol{\eta}$. We will assume a prior distribution in which θ and σ_u^2 are independent, with a uniform distribution with large support for θ, and a distribution for σ_u that is close to uniform on an interval assumed to contain the support of the full likelihood function for σ_u^2 with high probability. Gelman (2006) presents a thorough treatment of choosing a prior distribution of σ_u in the random effects model (1.1). He recommends using a uniform prior for σ_u for moderate to large values of n, but a half-Cauchy prior for smaller values of n (see, especially, Sections 3.2 and 5.2 of Gelman, 2006). The half-Cauchy prior is supported on $(0, \infty)$ and is given by

$$\pi(\sigma_u) \propto \left(1 + \left(\frac{\sigma_u}{A}\right)^2\right)^{-1}, \tag{2.4}$$

where A is a scale hyperparameter.

2.2 Unadjusted pairwise composite likelihood

Again, assume model (1.1), and assuming σ_e^2 known, let $\boldsymbol{\eta} = \left(\theta, \sigma_u^2\right)^{\mathrm{T}}$ be the vector of model parameters. We are interested in comparing the performance of the posterior distribution of $\boldsymbol{\eta}$ based on using the full likelihood or the pairwise likelihood, together with the adjusted posterior pairwise distribution to be described below.

To start, consider a simple situation where σ_u^2 also is assumed to be known and only θ is unknown. Let $\pi(\theta)$ be a prior density of θ. Then the posterior density of θ is

$$p_{\mathrm{FL}}\left(\theta \middle| \mathbf{y}(n)\right) \propto \pi(\theta) \prod_{i=1}^{n} f\left(\mathbf{y}_i; \theta\right), \tag{2.5}$$

where the subscript FL indicates that it is based on the full likelihood. In contrast, we consider

$$L_{i,\,\mathrm{PL}}(\theta) = \prod_{1 \le j < k \le m} L_{ijk}(\theta),$$

where $L_{ijk}(\theta) = \int f_{y|u}\left(y_{ij} \middle| u_i; \theta\right) f_{y|u}\left(y_{ik} \middle| u_i; \theta\right) f_u\left(u_i\right) du_i$, and then define

$$p_{\mathrm{PL}}\left(\theta \middle| \mathbf{y}(n)\right) \propto \pi(\theta) \prod_{i=1}^{n} L_{i,\,\mathrm{PL}}(\theta) \tag{2.6}$$

to be the "pairwise" posterior density of θ. We wish to compare the variances of θ derived from $p_{\mathrm{FL}}\left(\theta \middle| \mathbf{y}(n)\right)$ and $p_{\mathrm{PL}}\left(\theta \middle| \mathbf{y}(n)\right)$, shown in the following theorem, of which the derivations are straightforward.

Theorem: *Assume that $\pi(\theta)$ is a uniform prior. Then*

(a) $p_{\mathrm{FL}}\left(\theta \middle| \mathbf{y}(n)\right)$ *is a normal density with mean* $\bar{y}$ *and variance* $\frac{\sigma_e^2 + m\sigma_u^2}{mn}$;

(b) $p_{\mathrm{PL}}\left(\theta \middle| \mathbf{y}(n)\right)$ *is a normal density with mean* $\bar{y}$ *and variance* $\frac{\sigma_e^2 + 2\sigma_u^2}{m(m-1)n}$ *where* $\bar{y} = (mn)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} y_{ij}$.

The theorem shows that when m is greater than 2, the variance derived from the "pairwise" posterior density $p_{\mathrm{PL}}\left(\theta \middle| \mathbf{y}(n)\right)$ is smaller than that of the posterior density $p_{\mathrm{FL}}\left(\theta \middle| \mathbf{y}(n)\right)$. This finding is intuitively reasonable, because the pairwise likelihood is effectively taking all $m(m-1)/2$ pairs of observations within each cluster to be independent. It motivates us to examine an adjusted version of $p_{\mathrm{PL}}\left(\theta \middle| \mathbf{y}(n)\right)$, to be discussed in the sequel.

For the case where σ_u^2 is also unknown, it can be shown that a similar kind of adjustment is needed. Assuming independent uniform priors for θ and σ_u^2, it is straightforward to show that

$$p_{\mathrm{FL}}\left(\theta, \sigma_u^2 \middle| \mathbf{y}(n)\right) \propto \left|\boldsymbol{\Sigma}_m\right|^{-n/2} \exp\left[-0.5\,\mathrm{tr}\left(\boldsymbol{\Sigma}_m^{-1}\mathbf{S}_0\right)\right] \tag{2.7}$$

where $\mathbf{S}_0 = \sum_{i=1}^{n} (\mathbf{y}_i - \boldsymbol{\mu}_m)(\mathbf{y}_i - \boldsymbol{\mu}_m)^{\mathrm{T}}, \boldsymbol{\mu}_m = \theta\, \mathbf{1}_m, \boldsymbol{\Sigma}_m = \sigma_e^2 \mathbf{I}_m + \sigma_u^2 \mathbf{1}_m \mathbf{1}_m^{\mathrm{T}}, \mathbf{1}_m$ represents the $m \times 1$ unit vector, and $\mathbf{I}_m$ stands for the $m \times m$ identity matrix.

After some algebra the pairwise composite likelihood posterior (PL) can be shown to be

$$p_{\mathrm{PL}}\left(\theta, \sigma_u^2 \,\middle|\, \mathbf{y}(n)\right) \propto \left|\boldsymbol{\Sigma}_2\right|^{nm(m-1)/4} \exp\left[-0.5\,\mathrm{tr}\left(\boldsymbol{\Sigma}_2^{-1}\mathbf{S}_{0\mathrm{PL}}\right)\right] \tag{2.8}$$

where, with $\mathbf{z}_{ijk} = \left(y_{ij} - \theta,\ y_{ik} - \theta\right)^{\mathrm{T}}$,

$$\mathbf{S}_{0\mathrm{PL}} = \sum_{i=1}^{n} \sum_{j<k} \mathbf{z}_{ijk}\mathbf{z}_{ijk}^{\mathrm{T}}.$$

Note that $\boldsymbol{\Sigma}_2$ is defined in (2.7) with $m = 2$.

Assuming independent uniform priors for θ and σ_u^2, we consider the posterior density of σ_u^2 with θ integrated out. To assess the relative precisions of Bayesian inference in the two cases, we must use approximations because of the complexity of the two posterior densities. Specifically, we compare the curvature of the log posterior and the log pairwise posterior densities for σ_u^2 at their modes. The ratio of the latter to the former can be shown to be equal for large n to

$$\frac{2(m-1)\left(\sigma_e^2 + m\sigma_u^2\right)^2}{m\left(\sigma_e^2 + 2\sigma_u^2\right)^2},$$

implying that using the unadjusted pairwise posterior density for $m > 2$ would overestimate the precision of estimation of σ_u^2.

Thus, for both θ and σ_u^2 (or σ_u), basing an approximate log likelihood for Bayesian inference directly on the pairwise composite likelihood would lead to posterior intervals that are too narrow.

Note: In Section 3 the parameter vector $\boldsymbol{\eta}$ is set to be $\left(\theta, \sigma_u\right)^{\mathrm{T}}$ (with variance σ_u^2 replaced by standard deviation σ_u), and a half-Cauchy prior distribution is used for σ_u. However, the comparison of full and log pairwise posterior densities will remain similar under the appropriate transformations.

2.3 Curvature adjustment for the log pairwise likelihood

In this section we motivate the curvature adjustment for the log pairwise likelihood from the standpoint of estimating function theory, as presented, for example by Jørgensen and Knudsen (2004).

First, we note that if $\mathbf{X}$ has a q-variate normal distribution with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$, the logarithm of the multivariate density of $\mathbf{X}$ has form

$$-\frac{q}{2}\log(2\pi) - \frac{1}{2}\log\left|\boldsymbol{\Sigma}\right| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}). \tag{2.9}$$

The expression in (2.9) as a function of $\mathbf{x}$ has its maximum at $\boldsymbol{\mu}$ and curvature or second derivative matrix (Hessian) at the maximum equal to $-\boldsymbol{\Sigma}^{-1}$. Intuitively, this correspondence between the curvature of

the log density at the maximum and inverse of the covariance matrix can be expected to hold approximately for a multivariate density that is close to being normal.

Consider a model in which the distribution of the observation variable $\mathbf{Y}(n)$ depends on a vector parameter $\boldsymbol{\eta}$. Given an observation $\mathbf{Y}(n)=\mathbf{y}(n)$, the log likelihood is denoted $\ell(\boldsymbol{\eta};\mathbf{y}(n))=\log(f(\mathbf{y}(n);\boldsymbol{\eta}))$ where f is the density of $\mathbf{Y}(n)$. Under regularity conditions, (e.g., Lehmann, 1999, Chapter 7) the MLE $\hat{\boldsymbol{\eta}}$ is found by solving the system

$$\mathbf{s}(\boldsymbol{\eta};\mathbf{y}(n))=\mathbf{0}, \tag{2.10}$$

where $\mathbf{s}(\boldsymbol{\eta};\mathbf{y}(n))$ denotes the score function, the gradient of $\ell(\boldsymbol{\eta};\mathbf{y}(n))$. The system (2.10) is an unbiased (vector) estimating equation, and is optimally efficient, having minimal asymptotic variance-covariance matrix (in the sense of positive definite difference) among solutions of unbiased estimating equation systems. In regular cases (e.g., Lehmann, 1999, Chapter 7) the score function satisfies the *second Bartlett identity* (e.g., Lindsay, 1988):

$$\mathrm{Var}_{\boldsymbol{\eta}}\left[\mathbf{s}(\boldsymbol{\eta};\mathbf{y}(n))\right] = -E_{\boldsymbol{\eta}}\left[\nabla\mathbf{s}(\boldsymbol{\eta};\mathbf{y}(n))\right] = -E_{\boldsymbol{\eta}}\left[\nabla^2\ell(\boldsymbol{\eta};\mathbf{y}(n))\right], \tag{2.11}$$

where Var denotes a variance-covariance matrix, and ∇ represents a gradient. As well, asymptotically, through a Taylor series approximation of $\mathbf{s}(\hat{\boldsymbol{\eta}};\mathbf{y}(n))-\mathbf{s}(\boldsymbol{\eta};\mathbf{y}(n))=\mathbf{0}-\mathbf{s}(\boldsymbol{\eta};\mathbf{y}(\mathbf{n}))$, we have:

$$\hat{\boldsymbol{\eta}}-\boldsymbol{\eta} \simeq -\left[\nabla\mathbf{s}(\boldsymbol{\eta};\mathbf{y}(n))\right]^{-1}\mathbf{s}(\boldsymbol{\eta};\mathbf{y}(n)). \tag{2.12}$$

Thus, standard (frequentist) likelihood inference estimates the variance-covariance of $\hat{\boldsymbol{\eta}}$ as the reciprocal of the observed Fisher information matrix

$$\mathbf{I} = -\frac{\partial^2}{\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}^{\mathrm{T}}}\ell(\boldsymbol{\eta};\mathbf{y}(n))\Big|_{\hat{\boldsymbol{\eta}}} = -\nabla^2\ell(\boldsymbol{\eta};\mathbf{y}(n))\Big|_{\hat{\boldsymbol{\eta}}}, \tag{2.13}$$

which is the negative of the Hessian (curvature matrix) of the log likelihood function at its maximum.

In Bayesian inference, if $\pi(\boldsymbol{\eta})$ is a prior density for $\boldsymbol{\eta}$, the logarithm of the posterior density for $\boldsymbol{\eta}$ is

$$\log\pi(\boldsymbol{\eta}|\mathbf{y}(n)) = \log\pi(\boldsymbol{\eta}) + \ell(\boldsymbol{\eta};\mathbf{y}(n)) - K(\mathbf{y}(n)), \tag{2.14}$$

where

$$K(\mathbf{y}(n)) = \log\left\{\int\pi(\boldsymbol{\eta})f(\mathbf{y}(n);\boldsymbol{\eta})\,d\boldsymbol{\eta}\right\}.$$

If the prior density is flat in areas of appreciable likelihood, the posterior density of $\boldsymbol{\eta}$, which quantifies the inference about $\boldsymbol{\eta}$, approximates a density with mode at $\hat{\boldsymbol{\eta}}$ and the curvature of its logarithm equal to the negative of the Fisher information, making the posterior variance-covariance of $\hat{\boldsymbol{\eta}}$ approximately equal to the reciprocal of $\mathbf{I}$ in (2.13). Thus the Bayesian estimation of $\boldsymbol{\eta}$ is efficient in the frequentist sense; alternatively, the frequentist inference is close to the Bayesian inference.

Suppose that in the frequentist context, the score function is replaced by another estimating function $\mathbf{g}(\mathbf{y}(n); \boldsymbol{\eta})$ that is unbiased in the sense of having expectation 0. See, for example, Lindsay, Yi and Sun (2011). Then the estimator $\hat{\boldsymbol{\eta}}$ is no longer optimally efficient. However, it is consistent, and its variance can be estimated by the delta method, or linearization of the function $\mathbf{g}$. We might wish to think of treating $\mathbf{g}$ as a stand-in for a score vector, or as the gradient with respect to $\boldsymbol{\eta}$ of a substitute for the log likelihood function. In particular, composite likelihood equations might be thought of as stand-ins for score estimating equations.

A question is then whether a substitute for the log likelihood function having gradient $\mathbf{g}$ could play the role of the log likelihood in Bayesian inference, and lead to an approximately correct posterior when substituted into (2.14), and if not, whether there are principled ways in which we could correct it.

Thus, suppose we have an alternative to the score function, namely estimating function $\mathbf{g}(\mathbf{y}(n); \boldsymbol{\eta})$, that is unbiased for $\boldsymbol{\eta}$ in the sense of having

$$E_{\boldsymbol{\eta}}\left[\mathbf{g}(\mathbf{y}(n); \boldsymbol{\eta})\right] = \mathbf{0}.$$

Suppose the solution $\hat{\boldsymbol{\eta}}$ of the equation $\mathbf{g}(\mathbf{y}(n); \boldsymbol{\eta}) = \mathbf{0}$ maximizes a function $h(\mathbf{y}(n); \boldsymbol{\eta})$ which we would like to think of as an alternative to the log likelihood function; for example, $h(\mathbf{y}(n); \boldsymbol{\eta})$ could be a log pairwise composite likelihood function, and $\mathbf{g}(\mathbf{y}(n); \boldsymbol{\eta}) = \nabla h(\mathbf{y}(n); \boldsymbol{\eta})$. Then $h(\mathbf{y}(n); \boldsymbol{\eta})$ would be approximately equal to what the log posterior density would be if the prior were non-informative, and if we took $h(\mathbf{y}(n); \boldsymbol{\eta})$ to be a stand-in for the log likelihood function. The stand-in posterior variance-covariance of $\boldsymbol{\eta}$ would be approximately the inverse of the negative of the curvature matrix of $h(\mathbf{y}(n); \boldsymbol{\eta})$ at $\hat{\boldsymbol{\eta}}$. By estimating function theory (e.g., Heyde, 1997), using the same kind of Taylor series approximation as in (2.12), the frequentist variance-covariance of $\hat{\boldsymbol{\eta}}$ satisfies

$$\mathrm{Var}_{\boldsymbol{\eta}}\left(\hat{\boldsymbol{\eta}}^{\mathrm{T}}\right) \simeq \left\{E_{\boldsymbol{\eta}}\left[\nabla \mathbf{g}(\mathbf{y}(n); \boldsymbol{\eta})\right]\right\}^{-1} \mathrm{Var}_{\boldsymbol{\eta}}\left[\mathbf{g}(\mathbf{y}(n); \boldsymbol{\eta})\right]\left\{E_{\boldsymbol{\eta}}\left[\nabla \mathbf{g}(\mathbf{y}(n); \boldsymbol{\eta})\right]^{\mathrm{T}}\right\}^{-1}. \tag{2.15}$$

If $h(y; \boldsymbol{\eta})$ were the log pairwise composite likelihood function, we would have, in the notation of RCD,

$$\mathrm{Var}_{\boldsymbol{\eta}}\left(\hat{\boldsymbol{\eta}}^{\mathrm{T}}\right) \simeq \frac{1}{n}\left[\mathbf{H}(\boldsymbol{\eta}_0)\,\mathbf{J}(\boldsymbol{\eta}_0)^{-1}\,\mathbf{H}(\boldsymbol{\eta}_0)\right]^{-1}, \tag{2.16}$$

where $\boldsymbol{\eta}_0$ is the true value of $\boldsymbol{\eta}$, $n\mathbf{H}(\boldsymbol{\eta}_0)$ is minus the expectation of ∇h, and $n\mathbf{J}(\boldsymbol{\eta}_0)$ is equal to the variance-covariance matrix of $\mathbf{g}$, the gradient of h.

If $\mathbf{g}$ had the property (analogous to (2.11)) that

$$\mathrm{Var}_{\boldsymbol{\eta}}\left[\mathbf{g}(\mathbf{y}(n); \boldsymbol{\eta})\right] = -E_{\boldsymbol{\eta}}\left[\nabla \mathbf{g}(\mathbf{y}(n); \boldsymbol{\eta})\right], \tag{2.17}$$

so that $\mathbf{J}(\boldsymbol{\eta}_0) = -\mathbf{H}(\boldsymbol{\eta}_0)$, then the right-hand side of (2.15) or of (2.16) would be approximately the same as the stand-in posterior variance-covariance of $\boldsymbol{\eta}$.

The property (2.17) is called *information unbiasedness* of an estimating function (Lindsay, 1982). Given a $\mathbf{g}$ that does not satisfy (2.17), then to produce a $\mathbf{g}^*$ approximately satisfying (2.17), we could set

$$h^*\left(\mathbf{y}(n);\boldsymbol{\eta}\right) = h\left(\mathbf{y}(n);\hat{\boldsymbol{\eta}}+\mathbf{C}\left(\boldsymbol{\eta}-\hat{\boldsymbol{\eta}}\right)\right) = h\left(\mathbf{y}(n);\boldsymbol{\eta}^*\right) \tag{2.18}$$

for a constant matrix $\mathbf{C}$, so that the gradient of h^* is $\mathbf{C}^{\mathrm{T}}$ times the gradient of h, while the point estimate of $\boldsymbol{\eta}$ that maximizes h^*, and its approximate variance-covariance, are unchanged.

We want $\mathrm{Var}_{\boldsymbol{\eta}}\left(\mathbf{g}^*\right) = -E_{\boldsymbol{\eta}}\nabla\mathbf{g}^*$, and it can be shown that this is equivalent to

$$\mathbf{H}\left(\boldsymbol{\eta}_0\right)\mathbf{J}\left(\boldsymbol{\eta}_0\right)^{-1}\mathbf{H}\left(\boldsymbol{\eta}_0\right) = \mathbf{C}^{\mathrm{T}}\mathbf{H}\left(\boldsymbol{\eta}_0\right)\mathbf{C}, \tag{2.19}$$

which is a *curvature adjustment* like the one in RCD, who suggest taking the solution of (2.19) that sets $\mathbf{C} = \mathbf{M}^{-1}\mathbf{M}_A$, where $\mathbf{M}_A^{\mathrm{T}}\mathbf{M}_A = \mathbf{H}\left(\boldsymbol{\eta}_0\right)\mathbf{J}\left(\boldsymbol{\eta}_0\right)^{-1}\mathbf{H}\left(\boldsymbol{\eta}_0\right)$ and $\mathbf{M}^{\mathrm{T}}\mathbf{M} = \mathbf{H}\left(\boldsymbol{\eta}_0\right)$.

3. Simulation studies

3.1 Simulation design

Using simulation studies we have evaluated the performance of the proposed method, i.e., pairwise composite likelihood with a curvature adjustment, and compared it with using the full likelihood and the pairwise composite likelihood. We used the model in (1.1) to generate our data, i.e., for $i = 1,\ldots,n$ and $j = 1,\ldots,m$ we simulated values of Y_{ij} from

$$Y_{ij} = \theta + u_i + e_{ij}, \tag{3.1}$$

where $\theta = 1$, $u_i \overset{\text{iid}}{\sim} N\left(0,\sigma_u^2\right)$, and $e_{ij} \overset{\text{iid}}{\sim} N\left(0,\sigma_e^2\right)$. This is equivalent to having applied the superpopulation generation and sampling described in the paragraph surrounding (1.1).

Our first study, not included here, considered inference about θ with known σ_u and σ_e. It showed that using the pairwise composite likelihood for inference about θ badly overstated the precision, and that the curvature adjustment was successful. Thus, we proceeded to a more thorough study, considering inference for both θ and σ_u. To simplify we took $\sigma_e = 0.5$, and considered $n \in \{20, 40\}$ and $m \in \{5, 10\}$. For the half-Cauchy prior defined in (2.4) we took $A \in \{5, 10, 15\}$. There were 500 replicate data sets for each setting.

We considered three scenarios: (1) $\sigma_u \in \{0.1, 0.5\}$ and the half-Cauchy prior on σ_u; (2) Signal to Noise Ratio, $\mathrm{SNR} \in \{0.25, 0.75\}$ and the half-Cauchy prior on σ_u, where $\mathrm{SNR} = \sigma_u^2 / \left(\sigma_u^2 + \sigma_e^2\right)$; and (3) $\sigma_u \in \{0.1, 0.5\}$ and a uniform prior on σ_u. Throughout, we took a uniform prior on θ.

In Section 3.2 we describe the algorithms for the simulation studies.

3.2 Algorithms

As in Sections 2.1 and 2.2 define $\mathbf{y}(n) = \{\mathbf{y}_1,\ldots,\mathbf{y}_n\}$ with $\mathbf{y}_i = \left(y_{i1},\ldots,y_{im}\right)^{\mathrm{T}}$ and $\bar{y} = \sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij} / (mn)$. Further, $\boldsymbol{\eta}^{(t)}$ denotes the value of $\boldsymbol{\eta}$ at the tth iteration where $\boldsymbol{\eta} = \left(\theta,\sigma_u\right)^{\mathrm{T}}$. The full likelihood is

$$L_{\mathrm{FL}}\left(\theta,\sigma_u \mid \mathbf{y}(n)\right) \propto \left|\boldsymbol{\Sigma}_m\right|^{-n/2} \exp\left[-\frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Sigma}_m^{-1}\mathbf{S}_0\right)\right], \tag{3.2}$$

as in (2.7).

Using (3.2) together with the prior, $\pi(\boldsymbol{\eta})$, yields the posterior density,

$$p_{\mathrm{FL}}\left(\boldsymbol{\eta}\middle|\mathbf{y}(n)\right) \propto L_{\mathrm{FL}}\left(\boldsymbol{\eta}\middle|\mathbf{y}(n)\right)\pi(\boldsymbol{\eta}).$$

Sampling θ and σ_u is done in three steps:

Step 1. Sample $\theta^{(t)}$ from $p_{\mathrm{FL}}\left(\theta\middle|\mathbf{y}(n),\sigma_u^{(t-1)}\right)$ where

$$\theta \mid \left(\mathbf{y}(n),\sigma_u\right) \sim N\left(\bar{y},\frac{\sigma_e^2+m\sigma_u^2}{mn}\right).$$

We set the starting value, $\sigma_u^{(0)}$, to be the maximum likelihood estimate of σ_u.

Step 2. Use the Metropolis-Hastings (MH) algorithm to sample $\sigma_u^{(t)}$ from $p_{\mathrm{FL}}\left(\sigma_u \middle|\mathbf{y}(n),\theta^{(t)}\right)$. The latter is easily obtained from $p_{\mathrm{FL}}\left(\boldsymbol{\eta}\middle|\mathbf{y}(n)\right)$. Given $s>0$, the candidate σ_u, labelled σ_u^*, is sampled from the jumping distribution, $N\left(\sigma_u^{(t-1)},s^2\right)$. If $\sigma_u^*<0, \sigma_u^{(t)}=\sigma_u^{(t-1)}$. Otherwise, the procedure is standard with accept/reject ratio $p_{\mathrm{FL}}\left(\boldsymbol{\eta}_{\mathrm{FL}}^*\middle|\mathbf{y}(n)\right)\Big/p_{\mathrm{FL}}\left(\boldsymbol{\eta}_{\mathrm{FL}}^{(t-1)}\middle|\mathbf{y}(n)\right)$ where $\boldsymbol{\eta}_{\mathrm{FL}}^*=\left(\theta^{(t)},\sigma_u^*\right)^{\mathrm{T}}$ and $\boldsymbol{\eta}_{\mathrm{FL}}^{(t-1)}=\left(\theta^{(t)},\sigma_u^{(t-1)}\right)^{\mathrm{T}}$.

Step 3. Repeat Steps 1 and 2 for $K=1{,}000$ times with the first 200 samples used as the burn-in.

The pairwise composite likelihood (PL) is

$$L_{\mathrm{PL}}\left(\theta,\sigma_u \middle|\mathbf{y}(n)\right) \propto \left|\boldsymbol{\Sigma}_2\right|^{nm(m-1)/4} \exp\left[-\frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Sigma}_2^{-1}\mathbf{S}_{0\mathrm{PL}}\right)\right], \tag{3.3}$$

as in (2.8).

Using (3.3) together with the chosen prior, $\pi(\boldsymbol{\eta})$, yields the posterior density $p_{\mathrm{PL}}\left(\boldsymbol{\eta}\middle|\mathbf{y}(n)\right)$.

Sampling θ and σ_u is done in three steps:

Step 1. Sample $\theta^{(t)}$ from $p_{\mathrm{PL}}\left(\theta\middle|\mathbf{y}(n),\sigma_u^{(t-1)}\right)$ where

$$\theta \mid \left(\mathbf{y}(n),\sigma_u\right) \sim N\left(\bar{y},\frac{\sigma_e^2+2\sigma_u^2}{nm(m-1)}\right).$$

Step 2. Use the Metropolis-Hastings (MH) algorithm to sample $\sigma_u^{(t)}$ from $p_{\mathrm{PL}}\left(\sigma_u \middle|\mathbf{y}(n),\theta^{(t)}\right)$, as described in Step 2 above for the FL (substituting PL for FL in all formulas).

Step 3. Repeat Steps 1 and 2 for $K=1{,}000$ times with the first 200 samples used as the burn-in.

The final part is to obtain the (curvature) adjusted pairwise composite likelihood (APL), as described in Section 2.3. This derivation, based on the approach of RCD, exploits $\hat{\boldsymbol{\eta}}_{\mathrm{APL}}$, the estimated posterior means of θ and σ_u.

Step 1. Given (s_θ, s_σ) sample the candidate $\boldsymbol{\eta}^* = (\theta^*, \sigma_u^*)^{\mathrm{T}}$ from the bivariate normal jumping distribution, $N_2(\boldsymbol{\eta}^{(t-1)}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = \mathrm{diag}(s_\theta^2, s_\sigma^2)$. If $\sigma_u^* < 0, \boldsymbol{\eta}^{(t)} = \boldsymbol{\eta}^{(t-1)}$. Otherwise, go to Step 2.

Step 2. Define $\ell_{\mathrm{PL}}(\mathbf{y}(n) \mid \theta, \sigma_u)$ as the log pairwise composite likelihood obtained by taking the logarithm of (3.3), and $\ell_{\mathrm{PL}}(\mathbf{y}_i \mid \theta, \sigma_u)$ as the log pairwise composite likelihood corresponding to the data from cluster i, i.e., $\mathbf{y}_i$.

Step 3. Numerically obtain $\hat{\mathbf{H}} = \nabla^2 \ell_{\mathrm{PL}}(\mathbf{y}(n) \mid \hat{\theta}_{\mathrm{PL}}, \hat{\sigma}_{u\mathrm{PL}})$ and

$$\hat{\mathbf{J}} = \sum_{i=1}^{n} \left[\nabla \ell_{\mathrm{PL}}(\mathbf{y}_i \mid \hat{\theta}_{\mathrm{PL}}, \hat{\sigma}_{u\mathrm{PL}}) \left\{ \nabla \ell(\mathbf{y}_i \mid \hat{\theta}_{\mathrm{PL}}, \hat{\sigma}_{u\mathrm{PL}}) \right\}^{\mathrm{T}} \right],$$

where $\hat{\theta}_{\mathrm{PL}}$ and $\hat{\sigma}_{u\mathrm{PL}}$ are the estimated posterior means of θ and σ_u.

Step 4. Based on the approach of RCD, and using the singular value decomposition, we write $\hat{\mathbf{H}} = \mathbf{M}^{\mathrm{T}}\mathbf{M}$ and $\hat{\mathbf{H}}\hat{\mathbf{J}}^{-1}\hat{\mathbf{H}} = \mathbf{M}_A^{\mathrm{T}}\mathbf{M}_A$ for some matrices $\mathbf{M}$ and $\mathbf{M}_A$. Then define $\mathbf{C} = \mathbf{M}^{-1}\mathbf{M}_A$. In our case, $\mathbf{C}$ is a 2×2 matrix.

Step 5. From RCD the adjusted log pairwise composite likelihood, ℓ_{APL}, is

$$\ell_{\mathrm{APL}}(\mathbf{y}(n) \mid \boldsymbol{\eta}) = \ell_{\mathrm{PL}}(\mathbf{y}(n) \mid \boldsymbol{\eta}^*)$$

where

$$\boldsymbol{\eta}^* = \hat{\boldsymbol{\eta}}_{\mathrm{PL}} + \mathbf{C}(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}_{\mathrm{PL}}).$$

Step 6. Define the adjusted pairwise posterior density as

$$p_{\mathrm{APL}}(\boldsymbol{\eta} \mid \mathbf{y}(n)) \propto L_{\mathrm{APL}}(\mathbf{y}(n) \mid \boldsymbol{\eta})\, \pi(\theta, \sigma_u)$$

where $L_{\mathrm{APL}}(\mathbf{y}(n) \mid \boldsymbol{\eta}) = \exp(\ell_{\mathrm{APL}}(\mathbf{y}(n) \mid \boldsymbol{\eta}))$, the latter defined in Step 5.

Using the candidate value, $\boldsymbol{\eta}^*$, from Step 1 define the adjusted candidate value $\boldsymbol{\eta}_c^* = \hat{\boldsymbol{\eta}}_{\mathrm{PL}} + C(\boldsymbol{\eta}^* - \hat{\boldsymbol{\eta}}_{\mathrm{PL}})$. Then the accept/reject ratio is

$$p_{\mathrm{APL}}(\eta_c^* \mid \mathbf{y}(n)) \big/ p_{\mathrm{APL}}(\boldsymbol{\eta}^{(t)} \mid \mathbf{y}(n)).$$

The remaining steps are the standard ones for the Metropolis-Hastings algorithm.

3.3 Results from simulations

For each method (FL, PL, APL), each design parameter (m, n) and each prior distribution we summarized the simulation results using (a) the credible interval coverage rate in repeated sampling, and (b) the averages of the 0.025, 0.25, 0.50, 0.75 and 0.975 points of the posterior distributions of θ and σ_u.

There are also graphical summaries, i.e., *averaged posterior density* estimates for each of the posterior distributions, i.e., $p_{\mathrm{FL}}(\boldsymbol{\eta} \mid \mathbf{y}(n))$, $p_{\mathrm{PL}}(\boldsymbol{\eta} \mid \mathbf{y}(n))$ and $p_{\mathrm{APL}}(\boldsymbol{\eta} \mid \mathbf{y}(n))$. First, consider an interval, say, $[a, b]$, that supports most of the mass (e.g., 95%) of the posterior densities. Then divide it into $M = 50$ equally spaced subintervals with the cut points $a = c_0 < c_1 < \ldots < c_{M-1} < c_M = b$. For $t = 1, \ldots, T$, let

$\hat{f}_P^{(t)}(.)$ denote the estimate of the posterior density $f_P(.)$, derived from the tth simulation, where P stands for FL, PL or APL, and T is the number of simulations. Next define, for $r = 1, \ldots, M$,

$$\hat{f}_P(c_r) = \frac{1}{T}\sum_{t=1}^{T} \hat{f}_P^{(t)}(c_r).$$

Then a curve connecting the points $\{c_r, \hat{f}_P(c_r)\}$ for $a = c_0 < c_1 < \ldots < c_M = b$, is taken as the *averaged posterior density estimate for* $f_P(.)$.

Table 3.1 presents the coverage rates for θ and σ_u for $A = 15$, $n \in \{20, 40\}$, $m \in \{5, 10\}$, and $\sigma_u \in \{0.1, 0.289, 0.5, 0.866\}$. Figure 3.1 has the average posterior density estimates for θ and σ_u for $A = 15$, $\sigma_u \in \{0.1, 0.5\}$, $n = 40$, and $m = 10$. In both Table 3.1 and Figure 3.1 the summaries are given for the full likelihood (FL), pairwise composite likelihood (CL), and adjusted pairwise composite likelihood (APL).

Table 3.1
Coverage rates (in percent) for the 95% credible intervals of θ and σ_u with $A = 15$

		$\sigma_u = 0.1$		$\sigma_u = 0.289$		$\sigma_u = 0.5$		$\sigma_u = 0.866$	
		$n = 20$	$n = 40$	$n = 20$	$n = 40$	$n = 20$	$n = 40$	$n = 20$	$n = 40$
						θ			
$m = 5$	$\hat{\theta}_{\text{FL}}$	97.40	95.80	94.84	94.60	94.80	94.40	94.80	95.00
	$\hat{\theta}_{\text{PL}}$	68.20	66.60	58.45	58.40	53.60	51.40	50.00	50.20
	$\hat{\theta}_{\text{APL}}$	92.40	93.00	92.96	93.60	92.20	92.20	91.60	93.00
$m = 10$	$\hat{\theta}_{\text{FL}}$	94.80	95.00	95.00	94.00	94.80	94.20	95.00	93.80
	$\hat{\theta}_{\text{PL}}$	43.80	42.80	35.40	31.80	30.40	29.60	27.40	26.40
	$\hat{\theta}_{\text{APL}}$	90.60	91.80	92.20	93.40	92.80	92.60	91.80	93.00
$m = 5$	$\hat{\sigma}_{u,\text{FL}}$	97.20	99.00	91.55	95.40	93.00	94.80	92.60	95.00
	$\hat{\sigma}_{u,\text{PL}}$	92.80	85.60	59.62	61.80	52.40	54.20	46.20	48.20
	$\hat{\sigma}_{u,\text{APL}}$	88.40	83.40	86.85	92.20	84.40	91.20	82.00	89.60
$m = 10$	$\hat{\sigma}_{u,\text{FL}}$	99.00	97.20	93.60	92.80	93.80	93.80	93.00	93.60
	$\hat{\sigma}_{u,\text{PL}}$	63.60	56.80	33.40	38.00	27.00	29.60	24.40	26.60
	$\hat{\sigma}_{u,\text{APL}}$	82.80	84.40	85.20	89.00	80.80	86.60	79.00	87.00

The following summary includes the results for only the half-Cauchy prior with $A \in \{5, 10, 15\}$, $m \in \{5, 10\}$, $n \in \{20, 40\}$, and $\sigma_u \in \{0.1, 0.289, 0.5, 0.866\}$, the second and fourth values corresponding to SNR = 0.25 and SNR = 0.75, respectively. The results are similar for the three choices of A, and for the uniform prior.

Without any adjustment the coverages of PL differ substantially from the nominal 0.95. For example (Table 3.1), for $A = 15$, $n = 40$, $m = 10$, and σ_u = 0.5, the coverage for θ is less than 0.30. Considering

all values of the design parameters, the largest coverage is 0.70. In most cases, the coverage for θ is much less than 0.70.

With the curvature adjustment the coverage for θ is excellent. Of the 48 cases (three choices of A, two choices of m, two choices of n, four choices of σ_u), thirteen had coverage between 0.93 and 0.95, twenty-two between 0.92 and 0.93, eleven between 0.91 and 0.92, and two below 0.91, with the latter for $\sigma_u = 0.1$, $n = 20$, $m = 10$, and $A = 5$ and 15.

With the curvature adjustment the coverage for σ_u varies considerably, but there is, in almost all cases, a very large improvement in coverage relative to using the uncorrected pairwise composite likelihood.

The plots (Figure 3.1) show that for θ the posterior distribution corresponding to the adjusted likelihood is very close to the posterior distribution using the full likelihood. For σ_u there are differences between the posterior distributions corresponding to the full and adjusted likelihoods, most notably a shift to smaller values for the latter.

To investigate the effects of increasing m and n, consider the difference $\delta = C_{\text{FL}} - C_{\text{APL}}$ where C denotes coverage and FL and APL refer to the corresponding posterior distributions.

Overall with all m, n, A, and σ_u, for θ, δ decreases as n increases. For the larger values of σ_u, δ decreases as m increases, while for the smaller values of σ_u, δ tends to increase as m increases. Overall, for σ_u, δ decreases as n increases except in the case $\sigma_u = 0.1$, while δ increases as m increases.

The reason for the deterioration of the adjustment as m increases might be that the number of pairs per cluster is $m(m-1)/2$ and increases more rapidly, so that the pairwise likelihood quickly becomes more concentrated around its mode; the curvature adjustment may not suffice to compensate for a change in shape of the log pairwise composite likelihood, e.g., an increase in kurtosis.

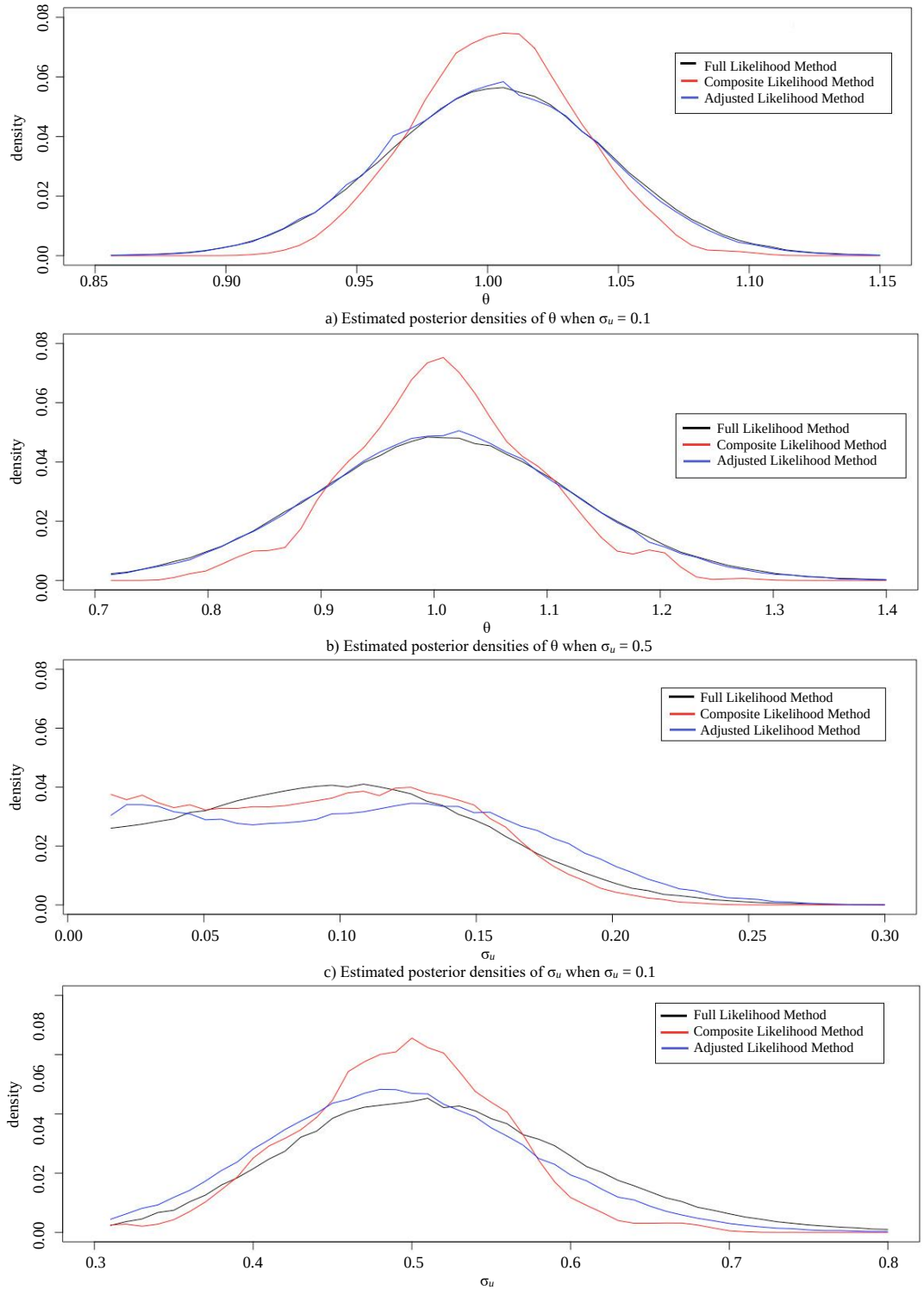
Table 3.2 presents one-sided non-coverage rates of the 95% Credible Intervals for θ and σ_u with $A = 15$, $n \in \{20, 40\}$, $m \in \{5, 10\}$, and $\sigma_u \in \{0.1, 0.289, 0.5, 0.866\}$. We observe the following:

i) For θ, the non-coverage for full likelihood intervals appears symmetric. The adjusted pairwise likelihood has undercoverage for θ, and except when σ_u is 0.1 the non-coverage is symmetric. A dependence of the coverage on m is seen only in the $\sigma_u = 0.1$ case.

ii) For σ_u, the full likelihood interval has non-coverage that is close to nominal and not very skewed, except in the case when $\sigma_u = 0.1$, where there is marked over-coverage. For $\sigma_u > 0.1$ and $m = 5$, coverage improves as n moves from 20 to 40, but for $\sigma_u > 0.1$ and $m = 10$, there is little difference in coverage for the two values of n.

iii) For σ_u, the adjusted pairwise likelihood has asymmetric non-coverage. Except in the case of $\sigma_u = 0.1$, the magnitude of the non-coverage tends to be similar on the left to that of the full likelihood, but much greater on the right, and the coverage improves as n moves from 20 to 40.

Figure 3.1 Estimated posterior densities of θ and σ_u using three methods when $A = 15$, $n = 40$, $m = 10$, and σ_u = (0.1, 0.5) using a half-Cauchy prior for σ_u.



a) Estimated posterior densities of θ when σ_u = 0.1

b) Estimated posterior densities of θ when σ_u = 0.5

c) Estimated posterior densities of σ_u when σ_u = 0.1

d) Estimated posterior densities of σ_u when σ_u = 0.5

Remembering that the adjusted log pairwise likelihood is not explicitly being constructed to approximate the log full likelihood, it does appear in Figure 3.1 that the adjusted log pairwise likelihood falls more quickly in the tails.

We also tried centering the curvature adjustment at the log pairwise posterior mode rather than the log pairwise posterior mean, and found that the under-coverage increased, though the asymmetry of coverage was less severe, for the resulting credible intervals.

Table 3.2
One-sided non-coverage rates (in percent) of the 95% Credible Intervals (CIs) of θ and σ_u with $A = 15$

		Non-CR-L	Non-CR-R	Non-CR-L	Non-CR-R	Non-CR-L	Non-CR-R	Non-CR-L	Non-CR-R
		$\sigma_u = 0.1$				$\sigma_u = 0.289$			
		$n = 20$		$n = 40$		$n = 20$		$n = 40$	
					θ				
$m = 5$	$\hat{\theta}_{\text{FL}}$	1.40	1.20	1.60	2.60	3.05	2.11	3.20	2.20
	$\hat{\theta}_{\text{PL}}$	16.60	15.20	16.40	17.00	21.13	20.42	20.40	21.20
	$\hat{\theta}_{\text{APL}}$	2.60	5.00	2.20	4.80	3.76	3.29	3.80	2.60
$m = 10$	$\hat{\theta}_{\text{FL}}$	2.80	2.40	2.40	2.60	3.00	2.00	3.20	2.80
	$\hat{\theta}_{\text{PL}}$	26.80	29.40	28.60	28.60	33.00	31.60	33.80	34.40
	$\hat{\theta}_{\text{APL}}$	3.60	5.80	4.60	3.60	4.00	3.80	3.80	2.80
					σ_u				
$m = 5$	$\hat{\sigma}_{u,\text{FL}}$	2.80	0.00	1.00	0.00	3.05	5.40	1.80	2.80
	$\hat{\sigma}_{u,\text{PL}}$	7.20	0.00	9.20	5.20	14.79	25.59	14.80	23.40
	$\hat{\sigma}_{u,\text{APL}}$	4.60	7.00	3.60	13.00	3.05	10.09	2.40	5.40
$m = 10$	$\hat{\sigma}_{u,\text{FL}}$	1.00	0.00	2.00	0.80	3.00	3.40	3.80	3.40
	$\hat{\sigma}_{u,\text{PL}}$	14.60	21.80	17.80	25.40	22.80	43.80	24.80	37.20
	$\hat{\sigma}_{u,\text{APL}}$	3.40	13.80	3.80	11.80	3.00	11.80	2.80	8.20
		$\sigma_u = 0.5$				$\sigma_u = 0.866$			
		$n = 20$		$n = 40$		$n = 20$		$n = 40$	
					θ				
$m = 5$	$\hat{\theta}_{\text{FL}}$	3.20	2.00	3.00	2.60	3.40	1.80	3.00	2.00
	$\hat{\theta}_{\text{PL}}$	24.40	22.00	24.60	24.00	26.40	23.60	25.80	24.00
	$\hat{\theta}_{\text{APL}}$	4.40	3.40	4.20	3.60	4.20	4.20	3.80	3.20
$m = 10$	$\hat{\theta}_{\text{FL}}$	3.00	2.20	3.40	2.40	3.00	2.00	3.40	2.80
	$\hat{\theta}_{\text{PL}}$	34.60	35.00	35.60	34.80	36.80	35.80	37.60	36.00
	$\hat{\theta}_{\text{APL}}$	4.00	3.20	4.20	3.20	4.80	3.40	3.40	3.60
					σ_u				
$m = 5$	$\hat{\sigma}_{u,\text{FL}}$	3.00	4.00	2.20	3.00	3.40	4.00	2.00	3.00
	$\hat{\sigma}_{u,\text{PL}}$	16.00	31.60	18.20	27.60	19.20	34.60	21.00	30.80
	$\hat{\sigma}_{u,\text{APL}}$	1.20	14.40	1.80	7.00	1.40	16.60	2.20	8.20
$m = 10$	$\hat{\sigma}_{u,\text{FL}}$	3.20	3.00	2.80	3.40	3.80	3.20	3.20	3.20
	$\hat{\sigma}_{u,\text{PL}}$	24.00	49.00	28.00	42.40	25.40	50.20	29.80	43.60
	$\hat{\sigma}_{u,\text{APL}}$	2.60	16.60	2.40	11.00	3.20	17.80	2.00	11.00

Note: Non-CR-L represents the left-side non-coverage rates (in percent) for the 95% CIs of θ and σ_u; Non-CR-R represents the right-side non-coverage rates (in percent) for the 95% CIs of θ and σ_u.

4. Extension to unequal probability sampling designs

An important extension of our setting is to a complex sampling framework, where frequentist parameter estimation through estimation of a population-level pairwise composite likelihood is now in fairly common use. RVH and YRL have shown that an approach based on applying a frequentist pairwise composite likelihood works well for estimating multilevel model variance components in the case of certain unequal probability sampling designs, and avoids the issue of inconsistency when the second stage sample sizes are small. The uncertainty estimation in this approach uses estimating function theory and may not require the adjustments we consider in this paper. However, it would be desirable to formulate a Bayesian counterpart of this method. If a Bayesian formulation were agreed upon, the results of our paper would predict a need for adjustment of the pseudo-log-pairwise-composite-likelihood to align it with an appropriate log full likelihood function.

Suppose that the purpose is still analytic, that the model for Y_{ij} is (1.1), and the objects of inference are the mean θ and the variance component σ_u^2 or its square root. The survey population has N first stage units with sizes M_i, $i = 1, \ldots, N$, and the first-stage sample consists of n of these, selected with an unequal probability sampling design. At the second stage, m_i elementary units are selected by simple random sampling from the ith first stage unit, if that unit has been sampled at the first-stage. If the sizes M_i and m_i and the sampling design probabilities $p(s)$ (where s runs through the two-stage subsets of the population satisfying the sample size specifications) do not depend on the u_i or e_{ij} values, the likelihood function can be taken to be of the form of (2.3), with m replaced by m_i, and the extension of our work is straightforward in principle. However, if the sizes or sampling design probabilities do depend on the values of u_i or e_{ij}, they will be informative about the parameters of interest. The sample-level likelihood function from the combination of multilevel model and sampling design may be ill-defined or intractable. From a Bayesian perspective we then need to consider what can reasonably substitute for the true likelihood, and how closely that substitute can be approximated by an adjusted pairwise composite likelihood. The answers may depend upon the preferred method of using the sampling design probabilities in inference, and there are several possibilities. Pursuing these possibilities would be a fruitful avenue for future research.

One method, with limited applicability, would be based on the approach of Léon-Novelo and Savitsky (2019). Assuming single stage Bernoulli sampling (so that the sampling probabilities are fully determined by the inclusion probabilities) they model the joint distribution of the outcome variable, Y, and the inclusion probability, π, using the model generating Y from $\mathbf{x}$ in the population and a model generating π from $\mathbf{x}$ and Y. To make computations feasible there are restrictions on the form of this model; see their Theorem 1 and, especially, the special case in their Section 2.1.

We can extend the model in Section 2.1 of Léon-Novelo and Savitsky (2019) to two-stage cluster sampling. A further extension, i.e., replacing the sampling density of Y with a pairwise composite likelihood analogous to the likelihood part of (2.6), can be made. Thus, subject to the limitations in

Theorem 1 of Léon-Novelo and Savitsky (2019), there are counterparts to the posterior densities, (2.5) and (2.6), that include the inclusion probabilities.

Another method, not fully Bayesian, but perhaps the most widely applicable extension of our approach, is to consider the population (census) log likelihood function ((2.5) and (2.6) of RVH) to be correct, and formulate a corresponding census log pairwise composite likelihood function as in our Section 2. We would then try to estimate the latter from the sample using sampling weights ((4.2) of RVH), and make adjustments such as appropriate weight normalization, or "scaling" as in Pfeffermann, Skinner, Holmes, Goldstein and Rasbash (1998), and curvature adjustments to the resulting estimated log pairwise composite likelihood function. This would produce a log pseudo-pairwise-likelihood function that could be used as an approximate log likelihood function in Bayesian inference. It would yield a Bayesian counterpart to the frequentist method put forward by RVH and YRL, and would extend the method of this paper to the unequal probability sampling situation.

We have obtained some preliminary details for this second approach. That is, if σ_u^2 is known, analytic expressions for the full likelihood and pairwise composite likelihood are available for θ at the census level. For the partial likelihood we alter (2.8) by taking σ_u fixed and add the weights w_i and $w_{jk|i}$ as in (4.2) of RVH. With a locally uniform prior for θ,

$$p_{\mathrm{PL}}\left(\theta \middle| \mathbf{y}(n)\right) \propto \exp\left\{-0.5 \sum_{i=1}^{n} \sum_{j<k} w_i w_{jk|i} \left(y_{ij} - \theta \quad y_{ik} - \theta\right) \Sigma_2^{-1} \left(y_{ij} - \theta \quad y_{ik} - \theta\right)^{\mathrm{T}}\right\}$$

where

$$\Sigma_2^{-1} = \begin{bmatrix} \sigma^{11} & \sigma^{12} \\ \sigma^{21} & \sigma^{22} \end{bmatrix}$$

with

$$\sigma^{11} = \sigma^{22} = \sigma_e^{-2}\left(1 - \frac{\sigma_u^2}{\sigma_e^2 + 2\sigma_u^2}\right)$$

and

$$\sigma^{12} = \sigma^{21} = -\frac{\sigma_u^2}{\sigma_e^2\left(\sigma_e^2 + 2\sigma_u^2\right)}.$$

After some algebra,

$$p_{\mathrm{PL}}\left(\theta \middle| \mathbf{y}(n)\right) \propto \exp\left\{-0.5 \frac{2\sum_{i=1}^{n}\sum_{j<k} w_i w_{jk|i}}{\sigma_e^2 + 2\sigma_u^2}\left[\theta - \frac{\sum_{i=1}^{n}\sum_{j<k} w_i w_{jk|i}\left(y_{ij} + y_{ik}\right)/2}{\sum_{i=1}^{n}\sum_{j<k} w_i w_{jk|i}}\right]^2\right\}.$$

Similarly, we alter (2.7) by taking σ_u fixed and adding the weights. With a locally uniform prior for θ,

$$p_{\mathrm{FL}}\left(\theta \mid \mathbf{y}(n)\right) \propto \exp\left\{-0.5\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{m} w_i w_{jk|i}\sigma^{jk}\left(y_{ij}-\theta\right)\left(y_{ik}-\theta\right)\right\}.$$

After some algebra,

$$p_{\mathrm{FL}}\left(\theta \mid \mathbf{y}(n)\right) \propto \exp\left\{-0.5\sum_{i=1}^{n} w_i\left\{\sum_{j=1}^{m} w_{j|i}a^{(1)} + \sum_{j\neq k} w_{jk|i}a^{(2)}\right\}\left(\theta-\hat{\theta}\right)^2\right\}$$

where

$$a^{(1)} = \sigma_e^{-2}\left(1-\frac{\sigma_u^2}{\sigma_e^2+m\sigma_u^2}\right),$$

$$a^{(2)} = -\frac{\sigma_u^2}{\sigma_e^2\left(\sigma_e^2+m\sigma_u^2\right)}$$

and

$$\hat{\theta} = \frac{\sum_{i=1}^{n} w_i\left[\sum_{j=1}^{m} a^{(1)}w_{j|i}y_{ij} + \sum_{j\neq k} a^{(2)}w_{jk|i}\left(y_{ij}+y_{ik}\right)/2\right]}{\sum_{i=1}^{n} w_i\left[\sum_{j=1}^{m} a^{(1)}w_{j|i} + \sum_{j\neq k} a^{(2)}w_{jk|i}\right]}.$$

Choice of the scaling of the weights will be important. To quantify the overstated precision in the log pairwise composite posterior a numerical evaluation may be required.

An advantage of pursuing extensions of this Bayesian approach further in future research would be that it is focused on inference for the model parameters rather than on finite population quantities, and thus it would not be necessary to bring third- or fourth-order inclusion probabilities into uncertainty estimation for σ_u^2 or σ_u.

5. Conclusion

There are well-known philosophical and foundational reasons for considering Bayesian approaches to survey sampling, and there is a long tradition of research in this area. See for example Sedransk (2008). There are also practical advantages. Using a Bayesian approach rather than a frequentist one relies much less on approximations, substituting computation for asymptotic expressions. In the context of random effects models, an important advantage is the ability to constrain the variance components to be non-negative in the prior distribution, without masking deficiencies in the data.

One example where Bayesian methods are used extensively is at the National Agricultural Statistical Service (NASS) of the US Department of Agriculture. At NASS, Bayesian methods are used to produce official statistics at the county and state levels for variables such as planted crop acreage and crop yield. Commonly, these inferences use several data sources. There is special attention to consistent estimation

across the hierarchy of geographical areas of interest for inference. See Nandram, Berg and Barboza (2014); Erciulescu, Cruze and Nandram (2020, 2019, 2018); and Cruze, Erciulescu, Nandram, Barboza and Young (2019) for additional details.

We have investigated a use of pairwise composite likelihood in Bayesian inference for survey data, in the sense of developing a posterior distribution for mean θ and standard deviation parameter σ_u of a simple random effects model. We have evaluated the posterior distribution in terms of the frequentist coverage properties of credible intervals for the parameters, and found them to work well for θ but not to be fully satisfactory for inference about σ_u for the settings considered. There would be corresponding implications for frequentist inference from the pairwise composite likelihood, treated as an approximate likelihood function. It is possible that better results might be obtainable through applying a suitable transformation to σ_u, and this is a subject of future research.

An ideal situation for the use of composite likelihood in Bayesian inference is one where (a) a model for generation of the data is fully specified, so that a true likelihood function exists, and (b) the true likelihood can be reasonably approximated by the composite likelihood, so that the corresponding posterior distributions agree well. For example, for Stoehr and Friel (2018) the motivation is the use for Bayesian inference of a pseudo-likelihood for data from a Gibbs random field. They establish identities that link the gradient and the Hessian of the log posterior for a parameter to moments of sufficient statistics of the random field, and use these to improve the ability of the log pairwise posterior density to approximate the log posterior density function. The curvature adjustment of RCD, upon which we have based our approach, instead adjusts the log pairwise composite likelihood so that its gradient (which we might call the "pairwise score vector") has the information-unbiasedness property that leads to credible intervals with frequentist coverage probabilities approximating nominal values. Intuitively, with the increase of the number n of clusters, m remaining fixed, this approximation should improve, and its computation does not require the use of properties of the likelihood itself. In this paper, we have used the availability of the full likelihood in the simple case to evaluate how closely Bayesian inference based on the adjusted pairwise composite likelihood resembles full Bayesian inference.

Acknowledgements

The research is partially supported by grants to Thompson and Yi from the Natural Sciences and Engineering Research Council of Canada (NSERC). Yi is Canada Research Chair in Data Science (Tier 1). Her research was undertaken, in part, thanks to funding from the Canada Research Chairs Program.

References

Cox, D.R., and Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 91, 729-737.

Cruze, N., Erciulescu, A., Nandram, B., Barboza, W. and Young, L. (2019). Producing official county-level agricultural estimates in the United States: Needs and challenges. *Statistical Science*, 34, 301-316.

Erciulescu, A., Cruze, N. and Nandram, B. (2018). Benchmarking a triplet of official estimates. *Environmental and Ecological Statistics*, 23, 523-547.

Erciulescu, A., Cruze, N. and Nandram, B. (2019). Model-based county level crop estimates incorporating auxiliary sources of information. *Journal of the Royal Statistical Society, Series A*, 182, 283-303.

Erciulescu, A., Cruze, N. and Nandram, B. (2020). Statistical challenges in combining survey statistics and auxiliary data to produce official statistics. *Journal of Official Statistics*, 36, 63-88.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 3, 515-533.

Heyde, C.C. (1997). *Quasi-Likelihood and its Application: A General Approach to Optimal Parameter Estimation.* New York: Springer-Verlag.

Jørgensen, B., and Knudsen, S.J. (2004). Parameter orthogonality and bias adjustment for estimating functions. *Scandinavian Journal of Statistics*, 31, 93-114.

Lehmann, E.L. (1999). *Elements of Large-Sample Theory*. New York: Springer-Verlag.

Léon-Novelo, L., and Savitsky, T. (2019). Fully Bayesian estimation under informative sampling. *Electronic Journal of Statistics*, 13, 1608-1645.

Lindsay, B.G. (1982). Conditional score functions: Some optimality results. *Biometrika*, 69, 505-512.

Lindsay, B.G. (1988). Composite likelihood methods. *Contemporary Mathematics*, 80, 220-239.

Lindsay, B.G., Yi, G.Y. and Sun, J. (2011). Issues and strategies in the selection of composite likelihoods. *Statistica Sinica*, 21, 71-105.

Nandram, B., Berg, E. and Barboza, W. (2014). A hierarchical Bayesian model for forecasting state-level corn yield. *Journal of Environmental and Ecological Statistics*, 21, 507-530.

Pfeffermann, D., Skinner, C.J., Holmes, D.J., Goldstein, H. and Rasbash, J. (1998). Weighting for unequal selection probabilities in multi-level models. *Journal of the Royal Statistical Society, Series B*, 60, 23-56.

Rao, J.N.K., Verret, F. and Hidiroglou, M.A. (2013). A weighted composite likelihood approach to inference for two-level models from survey data. *Survey Methodology*, 39, 2, 263-282. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2013002/article/11887-eng.pdf.

Rabe-Hesketh, S., and Skrondal, A. (2006). Multilevel modeling of complex survey data. *Journal of the Royal Statistical Society, Series A*, 169, 805-827.

Ribatet, M., Cooley, D. and Davison, A.C.D. (2012). Bayesian inference from composite likelihoods, with an application to spatial extremes. *Statistica Sinica*, 22, 813-845.

Sedransk, J. (2008). Assessing the value of Bayesian methods for inference about finite population quantities. *Journal of Official Statistics*, 24, 495-506.

Skinner, C.J., Holt, D. and Smith, T.F.M. (1989). *Analysis of Complex Surveys.* Wiley.

Stoehr, J., and Friel, N. (2018). Calibration of conditional composite likelihood for Bayesian inference on Gibbs random fields. *Artificial Intelligence and Statistics*, 921-929. arXiv:150201997v2.

Varin, C. (2008). On composite marginal likelihoods. *AStA Advances in Statistical Analysis*, 92, 1. https://doi.org/10.1007/s10182-008-0060-7 (accessed July 1, 2020).

Varin, C., Reid, N. and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 21, 5-24.

Yi, G.Y. (2017). Composite likelihood/pseudolikelihood. *Wiley StatsRef: Statistics Reference Online*. 1-14.

Yi, G.Y., Rao, J.N.K. and Li, H. (2016). A weighted composite likelihood approach for analysis of survey data under two-level models. *Statistica Sinica*, 26, 569-587.

Survey Methodology, June 2022
Vol. 48, No. 1, pp. 95-117
Statistics Canada, Catalogue No. 12-001-X

Non-response follow-up for business surveys

Elisabeth Neusy, Jean-François Beaumont, Wesley Yung, Mike Hidiroglou and David Haziza[1]

Abstract

In the last two decades, survey response rates have been steadily falling. In that context, it has become increasingly important for statistical agencies to develop and use methods that reduce the adverse effects of non-response on the accuracy of survey estimates. Follow-up of non-respondents may be an effective, albeit time and resource-intensive, remedy for non-response bias. We conducted a simulation study using real business survey data to shed some light on several questions about non-response follow-up. For instance, assuming a fixed non-response follow-up budget, what is the best way to select non-responding units to be followed up? How much effort should be dedicated to repeatedly following up non-respondents until a response is received? Should they all be followed up or a sample of them? If a sample is followed up, how should it be selected? We compared Monte Carlo relative biases and relative root mean square errors under different follow-up sampling designs, sample sizes and non-response scenarios. We also determined an expression for the minimum follow-up sample size required to expend the budget, on average, and showed that it maximizes the expected response rate. A main conclusion of our simulation experiment is that this sample size also appears to approximately minimize the bias and mean square error of the estimates.

Key Words: Non-response; Follow-up; Business surveys.

1. Introduction

Data collection research is a topic of interest amongst national statistical agencies looking to increase response rates and/or reduce data collection costs. With the high costs of collecting survey data, even a small increase in the efficiency of data collection procedures can translate into significant monetary savings. Given that response rates have declined over the past twenty years in both social and economic surveys, there has also been a growing concern over non-response bias.

In one of the first papers to discuss non-response, Hansen and Hurwitz (1946) proposed drawing a sub-sample of non-respondents, also called a non-response follow-up sample, to eliminate non-response bias. Their set-up was as follows: questionnaires were mailed out and after a certain period, a sample of non-respondents was followed up by personal interviewers to obtain their responses. They showed how the responses to the initial mail-out could be combined with those from the non-response follow-up sample to obtain an unbiased estimator of a population total or mean. They made the strong assumption that every unit of the follow-up sample responds. However, in today's environment, this assumption is not realistic as businesses and individuals are becoming increasingly reluctant to respond to surveys.

Much of the research published in the literature in the last 15 years has focused on adaptive collection designs, also called adaptive survey designs, responsive collection designs, responsive survey designs, or simply responsive designs. Groves and Heeringa (2006) defined a responsive survey design as one that uses paradata, or process data, to guide changes in the features of data collection to achieve higher quality

1. Elisabeth Neusy, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6. E-mail: elisabeth.neusy@statcan.gc.ca; Jean-François Beaumont, Wesley Yung and Mike Hidiroglou, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6. David Haziza, University of Ottawa, 150 Louis-Pasteur Private, Ottawa, ON, K1N 6N5.

estimates per unit cost. Beaumont, Bocci and Haziza (2014) noted that the literature on adaptive collection designs has mainly focussed on developing procedures that aim at reducing the non-response bias of an estimator that is not adjusted for non-response (see for example Schouten, Cobben and Bethlehem, 2009; and Peytchev, Riley, Rosen, Murphy and Lindblad, 2010). Beaumont et al. (2014) argued that any information (e.g., auxiliary data, paradata) that can be used during data collection to reduce non-response bias can also be used at the estimation stage. In other words, the non-response bias that can be removed at the collection stage through an adaptive collection procedure can also be removed at the estimation stage through appropriate non-response weight adjustments. They suggested that adaptive collection procedures, such as call prioritization, cannot reduce the non-response bias to a greater extent than a proper non-response weight adjustment. Limitations of adaptive collection procedures to reduce non-response bias and costs were also noted in the review paper by Tourangeau, Brick, Lohr and Li (2017).

So far, the literature on collection research has mostly targeted household surveys, and little has been reported on this subject for business surveys, two exceptions being Bosa, Godbout, Mills and Picard (2018) and Thompson, Kaputa and Bechtel (2018). Bosa et al. (2018) derived an item score that reflects the importance of following-up a particular sample unit and suggested an adaptative collection procedure using this score. Units with a large item score contribute the most to reducing the variance of point estimators. These units are given priority for expensive collection operations such as telephone follow-up. Thompson et al. (2018) considered sub-sampling of non-respondents and investigated the problem of sub-sample allocation subject to some constraints on the response rate and sample size in predetermined domains of interest.

Although business surveys typically use simple sampling designs, such as stratified simple random or Bernoulli sampling designs, they do possess certain features that can pose collection challenges. A distinctive feature is that business populations are highly skewed with a small percentage of businesses representing much of the economic activity. Consequently, business surveys usually include a take-all stratum where all units are selected with certainty, and take-some strata where the units are usually selected using simple random sampling without replacement or Bernoulli sampling. The take-all units correspond to large businesses. Failing to obtain a response from these large businesses could cause significantly biased estimates. As a result, all take-all units are typically followed up, and efforts are made to ensure their responses are received. The large businesses usually have staff (e.g., accountants) capable of responding to items on the questionnaire. On the other hand, small businesses may have to pay an outside accountant to obtain the requested information; this could be a contributing factor to non-response for such businesses. Another feature of business surveys is that collection is usually conducted in two steps. First, letters are sent to the sample units by postal service or by email, inviting them to complete an online electronic questionnaire. After a certain period of time, a follow-up of the non-responding units is conducted via computer-assisted-telephone interviews.

In this article, we focus on the take-some strata and attempt to respond to the following questions: (i) For a fixed budget for follow-up, how much effort should we dedicate to repeatedly following up

non-respondents until a response is received? (ii) Should we follow up all the non-respondents or select a sample of them? (iii) If we select a sample of non-respondents, what sampling designs would lead to more efficient estimators? To the best of our knowledge, determining an appropriate follow-up sample size and sampling design has not been investigated in the literature.

In the remainder of the paper, we present our investigations on non-response follow-up in the business survey context. The proposed follow-up strategy, which consists of a follow-up sampling design, data collection procedure, and estimator, is introduced in Section 2. In Section 3, we provide some theoretical properties of the proposed follow-up strategy. Section 4 describes a simulation study conducted to investigate the properties of the non-response-adjusted Hansen-Hurwitz estimator of a population total under different follow-up sampling designs and response scenarios. Finally, in Section 5, we summarize our main conclusions. Although we focus on business surveys, we believe that most of our conclusions also apply to social surveys.

2. Proposed follow-up strategy

Consider a finite population U of N units, partitioned into L strata, $U_1,\ldots,U_h,\ldots,U_L$, of size $N_1,\ldots,N_h,\ldots,N_L$, respectively, such that $U=\bigcup_{h=1}^{L}U_h$ and $N=\sum_{h=1}^{L}N_h$. We are interested in estimating the population total $Y=\sum_{h=1}^{L}\sum_{i\in U_h}y_{hi}$, where y_{hi} is the value of the variable of interest y for $i\in U_h$. From each stratum U_h, a sample s_{1h}, of size n_{1h}, is selected according to simple random sampling without replacement. The resulting total sample, $s_1=\bigcup_{h=1}^{L}s_{1h}$, is of size n_1. We denote by $\pi_{1hi}=n_{1h}/N_h$, the probability that unit $i\in U_h$ is selected in s_{1h}. The n_{1h} sampled units in stratum h are invited, either by post or email, to complete an online electronic questionnaire. We call this the "mail-out". If all sampled units respond to the mail-out, one could use the unbiased expansion estimator of Y, also called the full sample estimator:

$$\hat{Y}_{\text{FULL}}=\sum_{h=1}^{L}\sum_{i\in s_{1h}}w_{1hi}y_{hi}, \tag{2.1}$$

where $w_{1hi}=1/\pi_{1hi}$ denotes the design weight associated with $i\in s_{1h}$.

In practice, not all sampled units respond to the mail-out. Suppose that, after a certain period of time, n_{1hr} of the n_{1h} sampled units respond in stratum h. We denote the set of respondents in stratum h by s_{1hr}, and the response probability for unit $i\in s_{1h}$ by p_{1hi}. A sample of n_2 units, s_2, is then selected from the set of all non-respondents to the mail-out, $s_{1,\text{nr}}$. We denote by s_{2h}, the set of n_{2h} units selected for a follow-up in stratum h among the set of non-respondents to the mail-out in stratum h, $s_{1h,\text{nr}}$. We denote the probability that the mail-out non-respondent $i\in s_{1h,nr}$ is selected in the follow-up sample s_2 by π_{2hi}. We assume that this probability can be written as $\pi_{2hi}=n_2\pi^*_{2hi}$, where π^*_{2hi} does not depend on the follow-up sample size n_2 and satisfy the condition

$$\sum_{h=1}^{L}\sum_{i\in s_{1h,\text{nr}}}\pi^*_{2hi}=1. \tag{2.2}$$

This condition is satisfied for simple random sampling, stratified simple random sampling, with proportional or Neyman allocation, and probability proportional to size sampling.

Units of the sample s_2 are followed up via telephone. If all n_{2h} units respond to the follow-up, $h = 1, \ldots, L,$ the unbiased Hansen and Hurwitz (1946) estimator of the population total Y can be used:

$$\hat{Y}_{\mathrm{HH}} = \sum_{h=1}^{L} \sum_{i \in s_{1hr}} w_{1hi} y_{hi} + \sum_{h=1}^{L} \sum_{i \in s_{2h}} w_{1hi} w_{2hi} y_{hi}, \tag{2.3}$$

where $w_{2hi} = 1/\pi_{2hi}$ is the follow-up design weight of unit $i \in s_{2h}$. The objective of the sample s_2 is to estimate the unknown total $\sum_{h=1}^{L} \sum_{i \in s_{1h,nr}} w_{1hi} y_{hi}$. If a variable x strongly related to the variable of interest y is available before sample selection for all the mail-out non-respondents, it seems natural to use $w_{1hi} x_{hi}$ as an auxiliary variable for stratification or as a size measure for probability proportional to size sampling.

As pointed out by a reviewer, it is important to wait until mail-out data collection is closed before selecting the follow-up sample. If units respond to the mail-out after the follow-up sample has been selected, some decisions on how to handle these late respondents are required. If they are not discarded, it may be difficult to obtain an unbiased estimator like (2.3) without introducing model assumptions (see Beaumont, Bocci and Hidiroglou, 2014). This issue may also have implications on the length of the collection period.

As pointed out in the introduction, it is unlikely that all the follow-up sample units will respond. Suppose that after the end of the data collection period, n_{2hr} units have responded to the follow-up in stratum h. We denote by s_{2hr} the set of the n_{2hr} respondents in stratum h. We consider the non-response-adjusted version of the Hansen and Hurwitz (1946) estimator:

$$\hat{Y}_{\mathrm{HH-NA}} = \sum_{h=1}^{L} \sum_{i \in s_{1hr}} w_{1hi} y_{hi} + \sum_{h=1}^{L} \sum_{i \in s_{2hr}} w_{1hi} w_{2hi} a_{2hi} y_{hi}, \tag{2.4}$$

where a_{2hi} is a non-response weight adjustment. Under uniform non-response, a suitable weight adjustment is the inverse of the overall weighted response rate:

$$a_{2hi} = a_2 = \frac{\sum_{h=1}^{L} \sum_{j \in s_{2h}} w_{1hj} w_{2hj}}{\sum_{h=1}^{L} \sum_{j \in s_{2hr}} w_{1hj} w_{2hj}}, \quad i \in s_{2hr}, h = 1, \ldots, L. \tag{2.5}$$

A less restrictive assumption is uniform non-response within strata. Under this assumption, a suitable weight adjustment would be the inverse of the stratum weighted response rate:

$$a_{2hi} = a_{2h} = \frac{\sum_{j \in s_{2h}} w_{2hj}}{\sum_{j \in s_{2hr}} w_{2hj}}, \quad i \in s_{2hr}, h = 1, \ldots, L. \tag{2.6}$$

Note that the non-response weight adjustment (2.6) is computable only if $n_{2hr} > 0$ for all strata. Alternatively, unweighted versions of (2.5) and (2.6) could also be considered.

As mentioned earlier, follow-up of non-respondents who have been selected in s_2 is performed via telephone. In our proposed data collection procedure, a calling queue is first created by randomly ordering units in s_2. These units are then called sequentially until the queue is empty or the entire follow-up budget has been expended, whichever comes first. Each call attempt made to units in s_2 results in one of these three outcomes:

1. Response: A response is obtained from the unit. The unit is removed from the calling queue so that it does not get called again.
2. Final non-response: The unit is finalized as a non-respondent; it should not be called back again and is removed from the calling queue. The most common example of this outcome is a refusal to respond to the survey.
3. Still in progress: The unit is not finalized and needs to be called again; it is therefore returned to the end of the calling queue. An example of this outcome is an attempt where no contact is made or an attempt where an appointment is made for a callback.

The "response" and "final non-response" outcomes are both final outcomes, in the sense that the unit is removed from the calling queue and the collection process. This is in contrast to the "still-in-progress" outcome where the unit is returned to the calling queue so that it can be called again. A unit that completes the data collection process with an outcome of "response" or "final non-response" is said to be finalized or resolved, otherwise, it is said to be unresolved. There are two types of non-respondents after data collection: i) Finalized units with a "final non-response" outcome; and ii) Unresolved units. Both types of non-respondents are accounted for in estimation using the non-response-adjusted estimator (2.4).

We assume that, for a given sample unit, the outcomes of the call attempts are independent, and the probability associated with each of the three possible outcomes remains constant throughout the entire data collection period. For a given sampled unit $i \in s_{2h}, h = 1, \ldots, L,$ the probability of a "response" is denoted as $P_{2hi}^{(1)}$, the probability of a "final non-response" is denoted as $P_{2hi}^{(2)}$, and the probability of a "still-in-progress" outcome is denoted as $P_{2hi}^{(3)}$. In practice, the independence and constant probability assumptions may not hold exactly. The independence assumption is expected to be more plausible if the probabilities are conditional on strong predictors and if the time gap between two successive call attempts on the same unit is not too short. The constant probability assumption is not satisfied when the probabilities depend on predictors that can vary during data collection, such as the time of day or day of the week of the call attempt. Although it might be possible to extend our model to time-varying predictors, it would complicate our theoretical developments and simulation study. These assumptions are made throughout the paper to simplify our analyses. This is a limitation of our investigations that should be kept in mind when interpreting our results.

Multiple phone call attempts may be necessary to reach and resolve a unit. Data collection managers may wish to impose an upper limit on the number of call attempts that can be made to any follow-up sample unit. If a unit is still in progress after reaching the limit, it is removed from the calling queue and

remains unresolved at the end of data collection. Let K be that upper limit on the number of call attempts. Assuming each unresolved unit at the end of data collection always reaches the maximum number of attempts K, the probability that unit $i \in s_{1h,\mathrm{nr}}$ responds when selected in the sample s_2 can be written as $p_{2hi}(K) = \sum_{k=1}^{K} p_{2hik}$, where p_{2hik} is the probability that unit $i \in s_{1h,\mathrm{nr}}$ responds exactly at the kth attempt when selected in s_2. Under our assumptions, it is easy to see that $p_{2hik} = \left(P_{2hi}^{(3)}\right)^{k-1} P_{2hi}^{(1)}$. As a result, we have

$$\begin{aligned} p_{2hi}(K) &= \sum_{k=1}^{K} p_{2hik} \\ &= P_{2hi}^{(1)} \sum_{k=0}^{K-1} \left(P_{2hi}^{(3)}\right)^{k} \\ &= P_{2hi}^{(1)} \frac{1-\left(P_{2hi}^{(3)}\right)^{K}}{1-P_{2hi}^{(3)}} . \end{aligned} \tag{2.7}$$

In the next section, equation (2.7) will be used to determine an appropriate follow-up sample size.

3. Some theoretical properties of the proposed follow-up strategy

Let C be the total budget allocated for non-response follow-up, which could be defined in terms of monetary or time units. A cost is incurred for each call attempt and depends on the call outcome. We denote by $c^{(1)}$, $c^{(2)}$ and $c^{(3)}$, the cost per call attempt for a "response", "final non-response" and "still-in-progress" outcome, respectively. To simplify our derivations, we assume that these costs are the same for each sample unit and do not vary during data collection. Let $c_{hi} = \sum_{k=1}^{K} c_{hik}$ be the cost of either resolving unit $i \in s_{2h}$ or reaching the maximum number of call attempts for that unit, where c_{hik} is the cost of the kth call attempt for unit $i \in s_{2h}$. If a unit $i \in s_{2h}$ is resolved at the lth attempt, c_{hik} is defined to be zero for all $k > l$. Therefore, the cost c_{hik} is either zero, if unit $i \in s_{2h}$ has been resolved before the kth attempt, or $c^{(1)}$, $c^{(2)}$ or $c^{(3)}$, depending on the call outcome. For a given sample size n_2 and a fixed value of K, the total follow-up cost, $\sum_{h=1}^{L} \sum_{i \in s_{2h}} c_{hi}$, is a random variable when each sample unit is followed up until it is resolved or the maximum number of call attempts has been reached. Taking the expectation of the total cost with respect to the follow-up sampling design and non-response mechanism, conditionally on $s_{1,\mathrm{nr}}$, we obtain the expected follow-up cost:

$$\tilde{C}(n_2, K) = \sum_{h=1}^{L} \sum_{i \in s_{1h,\mathrm{nr}}} \pi_{2hi} \tilde{c}_{hi}(K), \tag{3.1}$$

where $\tilde{c}_{hi}(K) = \sum_{k=1}^{K} \tilde{c}_{hik}$ is the expected cost of either resolving unit $i \in s_{1h,\mathrm{nr}}$ or reaching the maximum number of call attempts, when that unit is selected in s_2, and $\tilde{c}_{hik}$ is the expected cost of the kth attempt, $k \le K$, for that unit. Given $c_{hik} \ne 0$ only if unit i has not been resolved before the kth attempt, it is easy to see that the expected cost $\tilde{c}_{hik}$ is

$$\tilde{c}_{hik} = \left(P_{2hi}^{(3)}\right)^{k-1} \left(c^{(1)} P_{2hi}^{(1)} + c^{(2)} P_{2hi}^{(2)} + c^{(3)} P_{2hi}^{(3)}\right).$$

The expected cost $\tilde{c}_{hi}(K)$ reduces to

$$\begin{aligned}\tilde{c}_{hi}(K) &= \sum_{k=1}^{K} \tilde{c}_{hik} \\ &= \left(c^{(1)}P_{2hi}^{(1)} + c^{(2)}P_{2hi}^{(2)} + c^{(3)}P_{2hi}^{(3)}\right)\sum_{k=0}^{K-1}\left(P_{2hi}^{(3)}\right)^{k} \\ &= \left(c^{(1)}P_{2hi}^{(1)} + c^{(2)}P_{2hi}^{(2)} + c^{(3)}P_{2hi}^{(3)}\right)\frac{1-\left(P_{2hi}^{(3)}\right)^{K}}{1-P_{2hi}^{(3)}}. \qquad (3.2)\end{aligned}$$

Using $\pi_{2hi} = n_2\pi_{2hi}^{*}$ along with condition (2.2), we can determine the follow-up sample size necessary to expend the budget C, on average, while ensuring each unit is resolved or has reached the maximum number of attempts, K. That is, we can determine the follow-up sample size such that the expected follow-up cost (3.1) is exactly equal to the budget C. This sample size is

$$n_2(C,K) = \frac{C}{\sum_{h=1}^{L}\sum_{i\in s_{1h,\mathrm{nr}}} \pi_{2hi}^{*}\tilde{c}_{hi}(K)}, \qquad (3.3)$$

where $\tilde{c}_{hi}(K)$ is given in (3.2). For a fixed budget C, the sample size $n_2(C,K)$ is inversely related to K and is a minimum when $K=\infty$; i.e., when there is no upper limit on the number of calls. This means that, for a fixed cost C, choosing a sample size larger than $n_2(C,\infty)$ has an effect similar to reducing the value of K, thereby increasing the expected number of unresolved units. Also, if a sample size smaller than $n_2(C,\infty)$ is chosen, the expected cost (3.1) is smaller than the budget C; i.e., on average, the budget is not entirely expended. The sample size $n_2(C,\infty)$ is thus the minimum sample size that expends the budget C, on average.

From the sample size $n_2(C,K)$ in (3.3), the expected number of respondents to the follow-up survey is

$$\begin{aligned}\tilde{n}_{2r}(C,K) &= \sum_{h=1}^{L}\sum_{i\in s_{1h,\mathrm{nr}}} \pi_{2hi}\,p_{2hi}(K) \\ &= C\,\frac{\sum_{h=1}^{L}\sum_{i\in s_{1h,\mathrm{nr}}} \pi_{2hi}^{*}\,p_{2hi}(K)}{\sum_{h=1}^{L}\sum_{i\in s_{1h,\mathrm{nr}}} \pi_{2hi}^{*}\tilde{c}_{hi}(K)}, \qquad (3.4)\end{aligned}$$

where $p_{2hi}(K)$ is given in (2.7), and the expected response rate is

$$\frac{\tilde{n}_{2r}(C,K)}{n_2(C,K)} = \sum_{h=1}^{L}\sum_{i\in s_{1h,\mathrm{nr}}} \pi_{2hi}^{*}\,p_{2hi}(K). \qquad (3.5)$$

From (2.7) and (3.5), we observe that the expected response rate does not depend on the budget C and decreases as K decreases. It was noted above that choosing a sample size larger than the minimum sample size $n_2(C,\infty)$, for a fixed cost C, has an effect similar to reducing the value of K. Consequently, choosing a sample size larger than $n_2(C,\infty)$ would also have the effect of reducing the expected response rate.

We can also obtain the expected number of resolved units in a way similar to (3.4) as

$$\tilde{n}_{2,\mathrm{res}}\left(C,K\right)=C\frac{\sum_{h=1}^{L}\sum_{i\in s_{1h,\mathrm{nr}}}\pi_{2hi}^{*}\left(1-\left(P_{2hi}^{(3)}\right)^{K}\right)}{\sum_{h=1}^{L}\sum_{i\in s_{1h,\mathrm{nr}}}\pi_{2hi}^{*}\tilde{c}_{hi}\left(K\right)}. \tag{3.6}$$

It can be easily seen that $\tilde{n}_{2,\mathrm{res}}\left(C,K\right)\le n_2\left(C,K\right)$ and that $\tilde{n}_{2,\mathrm{res}}\left(C,\infty\right)=n_2\left(C,\infty\right)$. If the follow-up sample size is chosen to be smaller than $n_2\left(C,\infty\right)$ then the expected cost $\sum_{h=1}^{L}\sum_{i\in s_{1h,\mathrm{nr}}}\pi_{2hi}\tilde{c}_{hi}\left(\infty\right)=C^{*}$, with $C^{*}<C$, and, from (3.4) and (3.6), both the expected number of respondents and resolved units decrease.

If the probability $P_{2hi}^{(3)}$ is very close to 1 for a few units $i\in s_{1h,\mathrm{nr}}$, $h=1,\ldots,L$, the minimum sample size $n_2\left(C,\infty\right)$ could become very small. In this situation, it may be appropriate to choose a finite value of K to avoid spending too large a portion of the budget on a few units. This would reduce the expected response rate, as noted above, and possibly increase the bias of estimates. However, using a finite value of K might also significantly increase the expected number of respondents and reduce the variance of estimates. Plots of the expected response rate and the expected number of respondents as a function of K may be useful to determine a suitable trade-off between the maximization of the expected response rate $\left(K=\infty\right)$ and the maximization of the expected number of respondents, which could be reached at a finite value of K. A small reduction of the expected response rate might be tolerated if it yields a significant increase in the expected number of respondents.

Under uniform follow-up response, we have: $P_{2hi}^{(1)}=P_2^{(1)}$, $P_{2hi}^{(2)}=P_2^{(2)}$ and $P_{2hi}^{(3)}=P_2^{(3)}$, for each unit $i\in s_{1h,\mathrm{nr}}$, $h=1,\ldots,L$. The follow-up sample size (3.3), the expected number of respondents (3.4), the expected response rate (3.5) and the expected number of resolved units (3.6) reduce to

$$n_2\left(C,K\right)=\frac{C}{\left(c^{(1)}P_2^{(1)}+c^{(2)}P_2^{(2)}+c^{(3)}P_2^{(3)}\right)}\frac{1-P_2^{(3)}}{1-\left(P_2^{(3)}\right)^{K}}, \tag{3.7}$$

$$\tilde{n}_{2r}\left(C,K\right)=\frac{C}{\left(c^{(1)}P_2^{(1)}+c^{(2)}P_2^{(2)}+c^{(3)}P_2^{(3)}\right)}P_2^{(1)}, \tag{3.8}$$

$$\frac{\tilde{n}_{2r}\left(C,K\right)}{n_2\left(C,K\right)}=P_2^{(1)}\frac{1-\left(P_2^{(3)}\right)^{K}}{1-P_2^{(3)}}, \tag{3.9}$$

and

$$\tilde{n}_{2,\mathrm{res}}\left(C,K\right)=\frac{C}{\left(c^{(1)}P_2^{(1)}+c^{(2)}P_2^{(2)}+c^{(3)}P_2^{(3)}\right)}\left(1-P_2^{(3)}\right), \tag{3.10}$$

respectively. It is worth pointing out that the expected number of respondents (3.8) and the expected number of resolved units (3.10) no longer depend on K. The expected number of resolved units,

$\tilde{n}_{2,\text{res}}(C, K)$, is therefore equal to the minimum sample size to expend the budget C, $n_2(C, \infty)$, for every value of K. As noted for the general expected response rate (3.5), the expected response rate (3.9) does not depend on the budget C and decreases as K decreases. Given the above observations, the value of K that maximizes both the expected response rate and the expected number of respondents is $K = \infty$ under uniform response, which leads to choosing the sample size $n_2(C, \infty)$.

The probabilities $P_{2hi}^{(1)}$, $P_{2hi}^{(2)}$ and $P_{2hi}^{(3)}$ are unknown. In practice, these probabilities must be replaced with estimates in the above expressions. Because they are needed before selecting the follow-up sample and collecting data, estimates of $P_{2hi}^{(1)}$, $P_{2hi}^{(2)}$ and $P_{2hi}^{(3)}$ could be obtained from previous survey data.

4. Simulation study

We conducted a simulation study to evaluate the properties of the non-response-adjusted estimator (2.4), $\hat{Y}_{\text{HH}-\text{NA}}$, under different response scenarios and follow-up sampling designs.

4.1 The simulation setup

Data used to create the sample s_1

The data used for the simulation study are sample data from an actual business survey: Statistics Canada's Monthly Survey of Food Services and Drinking Places (MSFSDP). As is typical for business surveys, the MSFSDP is stratified by province, industry and revenue (one take-all and one or more take-some strata within each province/industry combination). For greater detail on the MSFSDP, see Statistics Canada (2017). Each "Take All" stratum within a province/industry combination consists of the large and important businesses, which are usually all followed up. These units are excluded from the simulation study to focus on the follow-up strategy for the "Take some" strata. The set of sample units included in the simulation study is thus the original sample of 2,375 units selected in the $L = 63$ "Take some" strata.

Two variables are used for the simulation study: "Revenue" and "Sales". The first variable, Revenue, comes from the sampling frame (Statistics Canada's Business Register) and is present for all units selected in the MSFSDP sample. We use Revenue as an auxiliary variable, x, for sampling the non-respondents to the mail-out (see below). The second variable, Sales, is one of the variables collected by the survey; it is the variable of interest y. Both unit and item non-response are handled by imputation in the MSFSDP; thus Sales are available for all units in the simulation study and is imputed for 15% of the sample units. The correlation between Revenue and Sales is about 83% for both the respondent only data and the fully imputed data.

In our simulation experiments, the sample s_1 is not randomly generated multiple times from MSFSDP data. Instead, s_1 is fixed and consists of the set of all $n_1 = 2{,}375$ units in the original MSFSDP sample. The strata identifier, the design weight, the variable of interest y (Sales) and the auxiliary variable x (Revenue) for each unit of s_1 are taken from the MSFSDP sample file. Units with imputed y values are included in s_1, and imputed values are treated as observed values. This allows us to compute the full

sample estimate $\hat{Y}_{\text{FULL}}$ given in (2.1). This estimate is used as a benchmark to evaluate the properties of $\hat{Y}_{\text{HH-NA}}$ for different response scenarios and follow-up sampling designs, as detailed below.

Generation of the set $s_{1,\text{nr}}$ of mail-out non-respondents

Next, from s_1, response to the mail-out is generated independently from one unit to another using a Bernoulli distribution with probability p_{1hi}, $i \in s_{1h}$, $h = 1, \ldots, L$. Two response probability scenarios are considered:

1. Uniform: p_{1hi} = 50% for all sample units. Under this scenario, the expected number of non-respondents to the mail-out is 2,375/2 = 1,187.5.
2. Correlated to the variable of interest: p_{1hi} is determined using the logit function

$$\log\left(\frac{p_{1hi}}{1-p_{1hi}}\right) = -0.31 + 0.000004\, y_{hi}.$$

The constants -0.31 and 0.000004 are chosen by trial and error so that the expected number of non-respondents to the mail-out is again approximately half of the size of s_1. Note that the expected number of non-respondents to the mail-out can be written as $\sum_{h=1}^{L}\sum_{i\in s_{1h}}(1-p_{1hi})$. As a result, the constants are such that $\sum_{h=1}^{L}\sum_{i\in s_{1h}} p_{1hi} \approx 1{,}187.5$, where $p_{1hi} = \left[1+\exp\left(0.31-0.000004\, y_{hi}\right)\right]^{-1}$.

Selection of the follow-up sample s_2

The next step in the simulation is to select a follow-up sample s_2 from the set of mail-out non-respondents, $s_{1,\text{nr}}$, generated from one of the two response probability scenarios above. Five different sampling designs are considered for the selection of the follow-up sample:

1. Census of the mail-out non-respondents;
2. Simple Random Sampling (SRS) without replacement, ignoring the original stratification;
3. Stratified SRS without replacement using the original stratification, with sample allocation to strata proportional to the number of mail-out non-respondents;
4. Systematic sampling with probability proportional to Revenue, x_{hi}, ignoring the original stratification;
5. Systematic sampling with probability proportional to Revenue multiplied by the initial design weight, $w_{1hi}x_{hi}$, ignoring the original stratification.

Note that the size variables used for the two Probability Proportional to Size (PPS) sampling designs are trimmed from below the 5th percentile to remove zero-valued observations and some extremely small values that caused instability. On average, there are 1,188 non-respondents to the mail-out. For the first design, all non-respondents are followed up. For the remaining four designs, the follow-up sample sizes used for the simulation are chosen as 100, 200, 300, 400, 500, 700, and 900.

Generation of call outcomes

The outcomes of the telephone follow-up collection procedure are simulated at the call attempt level. For each sample unit $i \in s_{1h}$, $h = 1, \ldots, L$, the probabilities $P_{2hi}^{(1)}$, $P_{2hi}^{(2)}$ and $P_{2hi}^{(3)}$ for the three possible outcomes (see Section 2) are assigned before the start of the simulation and do not vary as data collection progresses. Two response scenarios are considered:

1. Uniform: $P_{2hi}^{(1)} = 25\%$, $P_{2hi}^{(2)} = 5\%$, and $P_{2hi}^{(3)} = 70\%$ for all units. These values were taken from Xie, Godbout, Youn and Lavallée (2011).
2. Correlated to the variable of interest: The probability of a "response" is based on the following logit function:

$$\log\left(\frac{P_{2hi}^{(1)}}{1-P_{2hi}^{(1)}}\right) = -1.29 + 0.000002\, y_{hi} + 0.3 z_{hi},$$

where z_{hi} is generated from the standard normal distribution. The constants -1.29, 0.000002 and 0.3 are chosen by trial and error so that the average of $P_{2hi}^{(1)}$ over all units in the sample s_1 is approximately 25%; i.e., $n_1^{-1} \sum_{h=1}^{L} \sum_{i \in s_{1h}} P_{2hi}^{(1)} \approx 0.25$, where $P_{2hi}^{(1)} = \left[1 + \exp(1.29 - 0.000002\, y_{hi} - 0.3 z_{hi})\right]^{-1}$. Note that the coefficient of correlation between the response probability $P_{2hi}^{(1)}$ and the variable of interest y_{hi} is 61%. The other two probabilities are defined as: $P_{2hi}^{(2)} = \frac{0.05}{0.75}\left(1 - P_{2hi}^{(1)}\right)$ and $P_{2hi}^{(3)} = \frac{0.70}{0.75}\left(1 - P_{2hi}^{(1)}\right)$. This ensures that $P_{2hi}^{(1)} + P_{2hi}^{(2)} + P_{2hi}^{(3)} = 1$.

For a given follow-up sample unit, the probabilities $P_{2hi}^{(1)}$, $P_{2hi}^{(2)}$ and $P_{2hi}^{(3)}$ are used to randomly generate the outcome of each call. After a call attempt, the unit returns to the end of the calling queue unless it is finalized and an outcome of "response" or "final non-response" is obtained. Outcomes are generated independently from one call to another. There is no explicit upper limit on the number of call attempts made to the same unit in our simulation study $(K = \infty)$.

Note that for the response scenario with varying response probabilities, the units that respond to the first call attempt are typically units with a higher response probability. As a result, the units that remain in the calling queue for the second attempt tend to be units with a lower response probability. It follows that the proportion of units that respond in the second attempt tends to be lower than in the first attempt. Similarly, the proportion of units that respond in the third attempt tends to be lower than in the second attempt, and so on. The proportion of units that respond decreases with each call attempt, as the units that remain in the calling queue are those that are harder to reach. Therefore, estimates may suffer from substantial bias if data collection ends prematurely, and if those that are harder to reach tend to have y-values larger or smaller than the other sample units.

The total budget for follow-up is fixed at 3,000 units (monetary or time units) in our study. A cost is charged for each call attempt. The amount charged depends on the outcome of the attempt: a "response" outcome has a cost of 5 units $(c^{(1)} = 5)$, a "final non-response" outcome has a cost of 2 units $(c^{(2)} = 2)$, and a "still-in-progress" outcome has a cost of 1 unit $(c^{(3)} = 1)$. The collection ends when the budget runs

out, or when there are no more cases left in the calling queue (i.e., all units are resolved), whichever occurs first. The cost values and budget have been chosen somewhat arbitrarily as they are survey-specific. However, we ensured that $c^{(1)} > c^{(2)} > c^{(3)}$ as this relation is generally expected to hold in telephone surveys.

Monte Carlo measures

The generation of responses to the mail-out, the selection of the follow-up sample and the generation of responses to the follow-up are repeated independently $R = 1{,}000$ times for each combination of mail-out response scenario, follow-up sampling design and follow-up response scenario described above. The non-response-adjusted estimator (2.4), $\hat{Y}_{\mathrm{HH-NA}}$, is computed for each replicate. The non-response weight adjustments a_{2hi} are computed using (2.5) as the inverse of the overall weighted response rate. We use $a_{2hi} = a_2$, given in (2.5), rather than $a_{2hi} = a_{2h}$, given in (2.6), to avoid a few cases where some of the sets s_{2hr} are empty, which would lead to infinite values of a_{2h}. The non-response weight adjustment (2.5) can be viewed as an extreme form of collapsing. Less extreme collapsing could be applied in practice and might show better properties. We choose (2.5) in this simulation study for its simplicity.

Using the 1,000 replicates of $\hat{Y}_{\mathrm{HH-NA}}$, the Monte Carlo Relative Bias (RB) and Relative Root Mean Square Error (RRMSE) of $\hat{Y}_{\mathrm{HH-NA}}$ are computed as

$$\mathrm{RB} = \frac{1}{R}\sum_{r=1}^{R} E_r \times 100\% \quad \text{and} \quad \mathrm{RRMSE} = \sqrt{\frac{1}{R}\sum_{r=1}^{R} E_r^2} \times 100\%,$$

where $E_r = \left(\hat{Y}^r_{\mathrm{HH-NA}} - \hat{Y}_{\mathrm{FULL}}\right) \Big/ \hat{Y}_{\mathrm{FULL}}$ is the relative error for the rth simulation replicate, and $\hat{Y}^r_{\mathrm{HH-NA}}$ is the non-response-adjusted Hansen-Hurwitz estimator for the rth replicate, $r = 1, \ldots, 1{,}000$.

As pointed out above, the initial sample s_1 is fixed for each of the 1,000 replicates to focus on the mail-out and follow-up response mechanisms and the follow-up sampling design. While it could have been possible to create an artificial population and draw a different initial sample at each replicate, it was felt that this additional complexity would not change our main conclusions, except for systematically increasing the variance of $\hat{Y}_{\mathrm{HH-NA}}$. Our simulation setup has also the advantage of being conditional on real sample data.

4.2 Simulation results

In this section, we discuss the simulation results for four scenarios of mail-out and follow-up response:

1. The response probability is uniform for both the mail-out and the follow-up. This serves as a baseline scenario with which to compare the other scenarios.
2. The response probability is correlated to Sales for the mail-out and uniform for the follow-up.
3. The response probability is uniform for the mail-out and correlated to Sales for the follow-up.

4. The response probability is correlated to Sales for both the mail-out and the follow-up. This scenario is probably the most realistic.

Response Scenario 1: Uniform response probability for both the mail-out and the follow-up

Figure 4.1 shows the relative bias versus the follow-up sample size for the five sampling designs. Figure 4.2 shows the RRMSE versus the follow-up sample size. Note that the results for the follow-up of all mail-out non-respondents are given by the last point on the figures (i.e., a sample size of 1,188).

Figure 4.1 Relative bias versus follow-up sample size for Scenario 1.

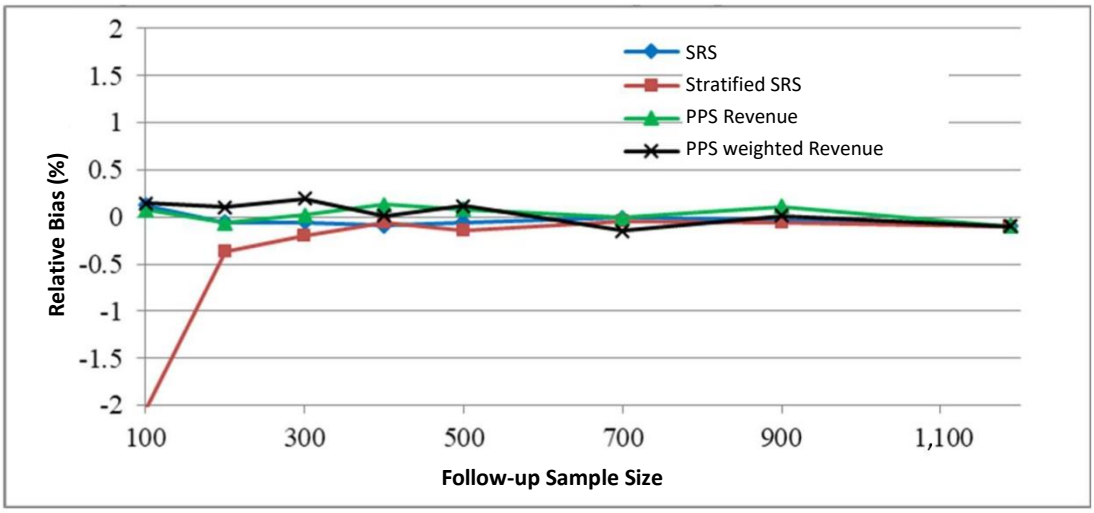
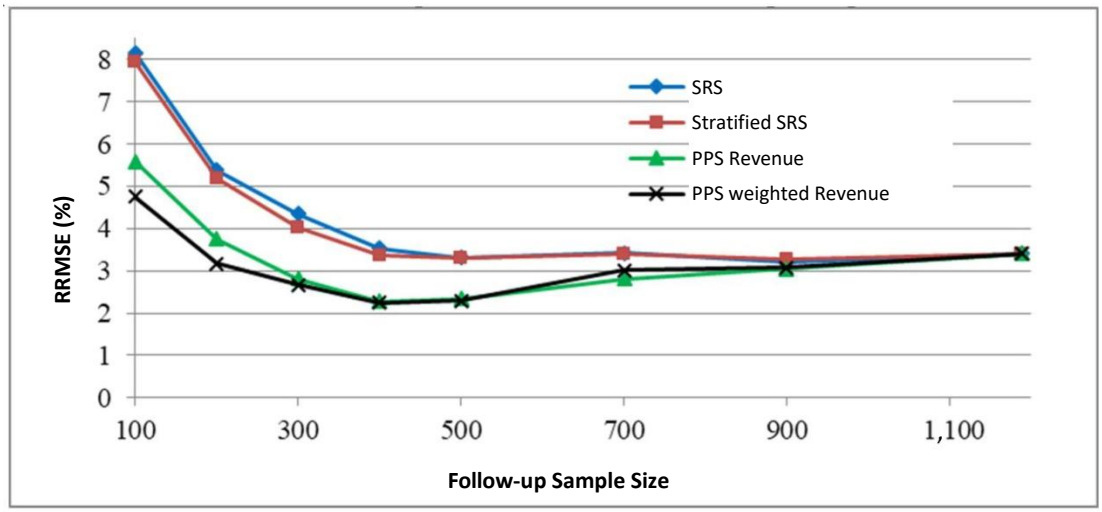


Figure 4.2 Relative root mean square error versus follow-up sample size for Scenario 1.



The following observations can be made by examining Figures 4.1 and 4.2:

- The RB is approximately zero for all follow-up sample sizes and designs. The only exception is stratified SRS with a follow-up sample size of 100. The proportional allocation strategy for the follow-up sample does not ensure that at least one unit is selected from each stratum. Therefore, for smaller follow-up sample sizes (e.g., 100), some strata end up with no follow-up sample although they may contain mail-out non-respondents. This causes a negative bias for the estimation of a population total.
- As the sample size increases from 100 to 400, the RRMSE decreases for all designs. This can be explained by an increase of the average number of respondents as the sample size increases (not shown in the figures).
- For sample sizes greater than 400, the RRMSE remains roughly constant for the SRS and stratified SRS designs. For those sample sizes, the average number of respondents remains roughly constant. This is consistent with equation (3.8). It indicates that, under uniform response to the follow-up, the expected number of respondents does not vary with K, and thus with the follow-up sample size, provided the budget is expended.
- The PPS designs seem to be more efficient than the SRS and stratified SRS designs. However, for sample sizes greater than 400, the gains in efficiency diminish as the sample size increases.

Response Scenario 2: Response probability correlated to Sales for the mail-out and uniform for the follow-up

Figures 4.3 and 4.4 show the relative bias and the RRMSE for Scenario 2, respectively.

Figure 4.3 Relative bias versus follow-up sample size for Scenario 2.

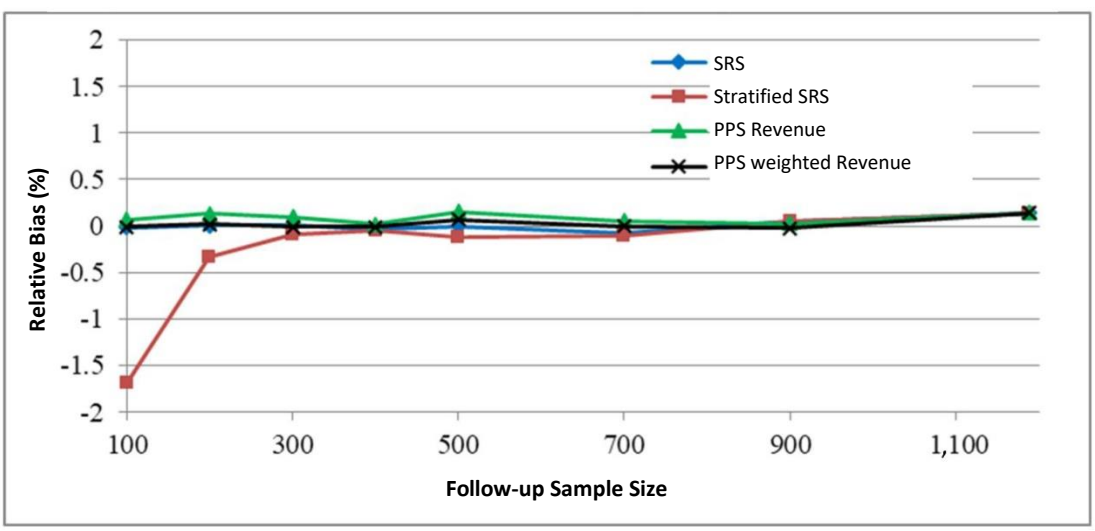
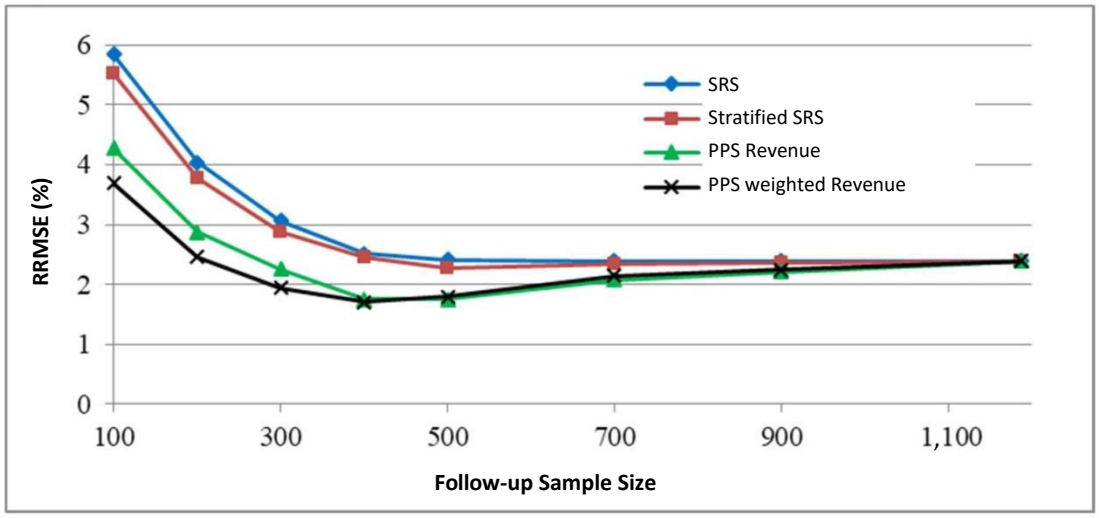


Figure 4.4 Relative root mean square error versus follow-up sample size for Scenario 2.



The following observations can be made by examining Figures 4.3 and 4.4:

- The results show that if the mail-out response probability is correlated to Sales, but the follow-up response probability is uniform, the bias can be nearly eliminated through the follow-up sampling design. This can be explained by observing that the Hansen and Hurwitz (1946) estimator (2.3) is unbiased for any mail-out response mechanism.
- The observations given for Scenario 1 apply to Scenario 2 as well.

Response Scenario 3: Response probability uniform for the mail-out and correlated to Sales for the follow-up

Figures 4.5 and 4.6 show the relative bias and the RRMSE for Scenario 3, respectively.

Figure 4.5 Relative bias versus follow-up sample size for Scenario 3.

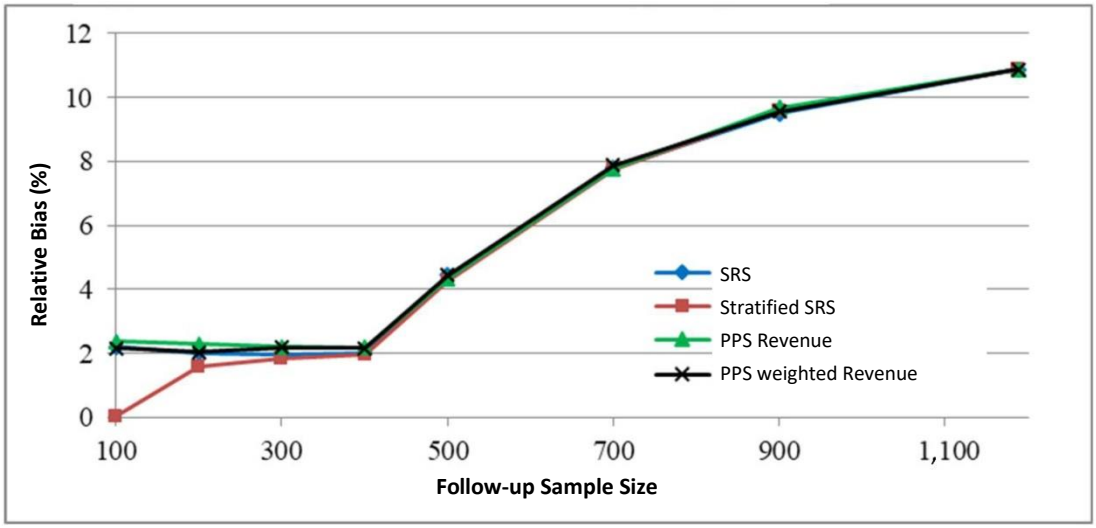
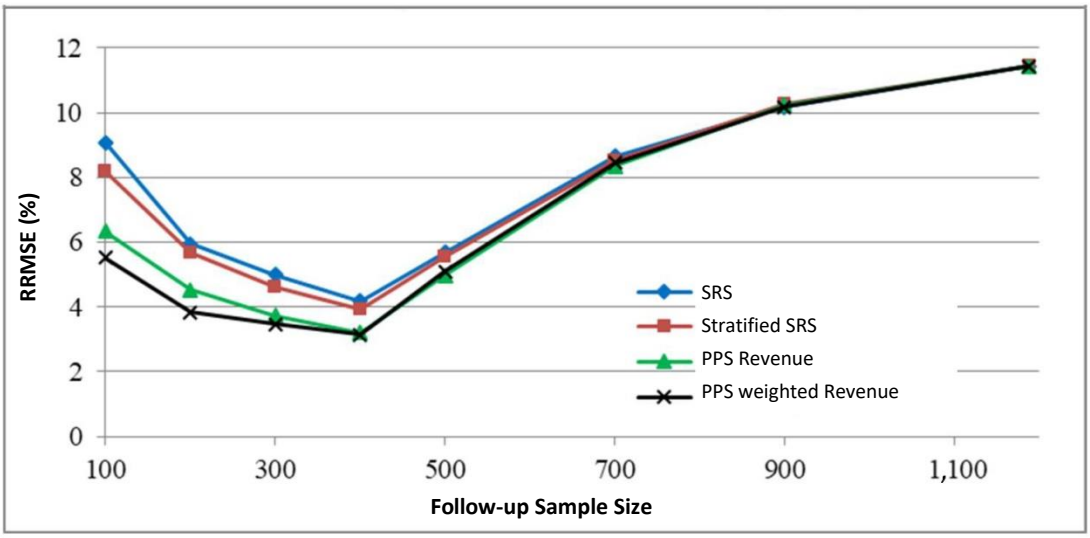


Figure 4.6 Relative root mean square error versus follow-up sample size for Scenario 3.



The following observations can be made by examining Figures 4.5 and 4.6:

- The RB is lowest for sample sizes less than or equal to 400, where we observed that all the units were finalized before the budget ran out. The lower RB for stratified SRS with a follow-up sample size of 100 is due to strata with no follow-up sample (see Response Scenario 1).
- The RRMSE is minimized for a sample size of 400.
- For sample sizes greater than 400, both RB and RRMSE increase as the sample size increases. For those sample sizes, we observed a diminution of the average response rate as the sample size increases (see the discussion below equation (3.5) for a theoretical justification). This explains the increase of RB and RRMSE as the sample size increases.
- The PPS designs seem again to be more efficient than the SRS and stratified SRS designs. However, for sample sizes greater than 400, the gains in efficiency diminish as the sample size increases.

Response Scenario 4: Response probability correlated to Sales for both the mail-out and the follow-up

Figures 4.7 and 4.8 show the relative bias and the RRMSE for Scenario 4, respectively.

Figures 4.7 and 4.8 are similar to Figures 4.5 and 4.6. The observations given for Scenario 3 apply to Scenario 4 as well.

Figure 4.7 Relative bias versus follow-up sample size for Scenario 4.

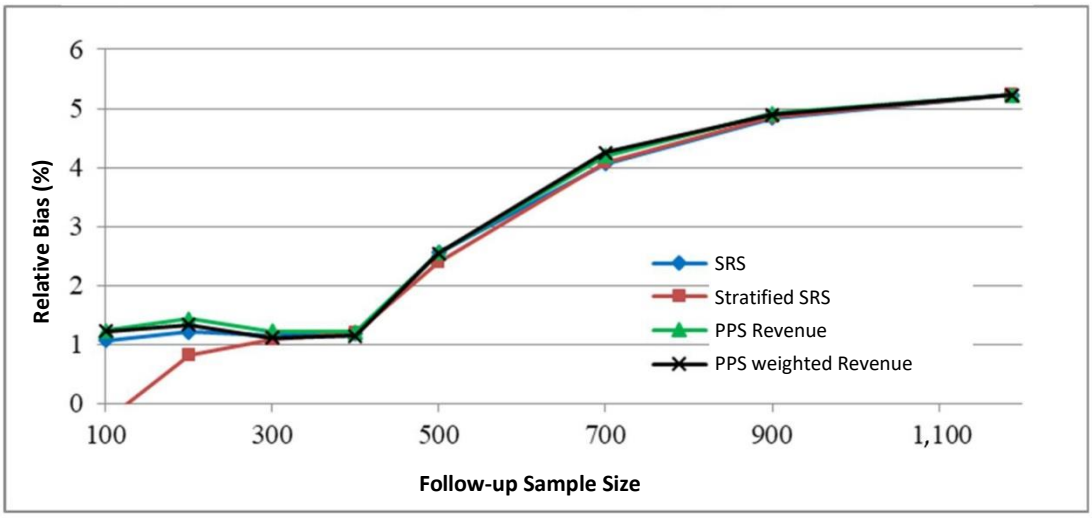
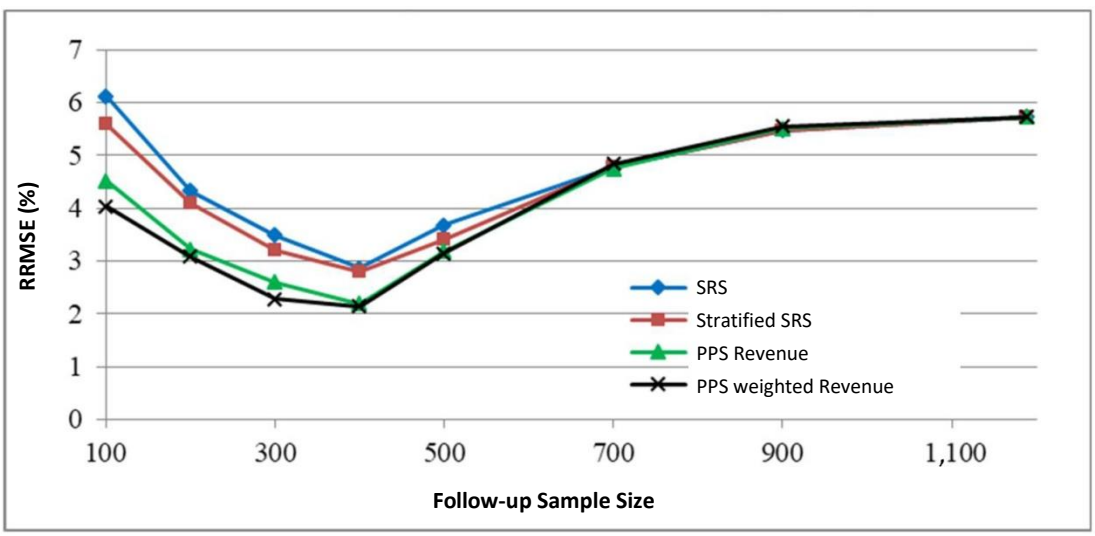


Figure 4.8 Relative root mean square error versus follow-up sample size for Scenario 4.



4.3 Remarks on the simulation results

We observed that for follow-up sample sizes smaller than or equal to 400, and for all sampling designs and response scenarios, all the units were finalized with an outcome of "response" or "final non-response" before the budget was exhausted, except for two simulation replicates. As a result, the follow-up response rate remained roughly constant whereas the number of respondents increased as the follow-up sample size increased from 100 to 400, reducing the variance and mean square error of the estimator $\hat{Y}_{\mathrm{HH-NA}}$.

For sample sizes of 500 or over, the follow-up budget always ran out before all the units were finalized. As the follow-up sample size increased, the number of respondents and finalized units remained

roughly constant. On average, between 430 and 445 cases were finalized at the end of data collection depending on the sampling design and response scenario; the other units were left in the calling queue with an outcome of "still-in-progress". It thus appears that the follow-up budget used for the simulation study was just large enough to finalize around 440 units for sample sizes greater than or equal to 500. Given that the number of respondents remained roughly constant as the sample size increased, the response rate decreased. The reduction of the response rate can be explained by a smaller average number of call attempts per sample unit as the follow-up sample size increases. This has the undesirable consequence of increasing the bias and mean square error of $\hat{Y}_{\mathrm{HH-NA}}$ for the non-uniform follow-up response mechanism.

From Figures 4.2, 4.4, 4.6 and 4.8, we also observe that the RRMSE reaches a minimum for a sample size of 400 or 500 depending on the response scenario and sampling design. The sample size that minimizes the RRMSE seems to correspond roughly to the minimum sample size that expends the follow-up budget on average. As discussed above, a smaller sample size increases the variance of $\hat{Y}_{\mathrm{HH-NA}}$, due to a smaller number of respondents, whereas a larger sample size may increase the bias due to a reduced response rate. The minimum sample size to expend the follow-up budget appears to be the same as the expected number of resolved units, which was around 440 in our simulation study for sample sizes of 500 or above.

The theory developed in Section 3 supports the above empirical observations for uniform response to the follow-up. Table 4.1 provides values of the sample size (3.7), the expected number of respondents (3.8), the expected response rate (3.9), and the expected number of resolved units (3.10) for different values of K, and for the values of C, $c^{(1)}$, $c^{(2)}$, $c^{(3)}$, $P_2^{(1)}$, $P_2^{(2)}$ and $P_2^{(3)}$ used in the simulation study: $C = 3{,}000$, $c^{(1)} = 5$, $c^{(2)} = 2$, $c^{(3)} = 1$, $P_2^{(1)} = 0.25$, $P_2^{(2)} = 0.05$ and $P_2^{(3)} = 0.70$. The minimum sample size $n_2(C, \infty)$ and the expected number of resolved units $\tilde{n}_{2,\mathrm{res}}(C, K)$ are equal to 439; this agrees with the simulation results.

As shown in Table 4.1, a small value of K may reduce significantly the expected response rate whereas the expected number of respondents does not vary with K provided the budget is expended. Therefore, under uniform response to the follow-up, there does not seem to be any advantage to using a follow-up sample size larger than $n_2(C, \infty)$, the minimum sample size to expend the budget on average, which is 439 in this scenario. This choice maximizes the expected response rate without reducing the expected number of respondents. Under moderate departure from uniform response, choosing a sample size close to $n_2(C, \infty)$ (or a large value of K) would ensure the non-response bias is better controlled.

Our simulation results indicate that the conclusions drawn from Table 4.1 hold approximately for non-uniform response to the follow-up. In particular, the minimum sample size that expends the budget was close to 439 and the expected number of respondents and resolved units stayed roughly constant when the follow-up sample size increased. As a result, incorrectly assuming uniform response when it is not uniform leads to an appropriate sample size in our simulation setup. Another conclusion of our simulation study is that choosing a follow-up sample size close to $n_2(C, \infty)$ appears to minimize both the

non-response bias and mean square error of $\hat{Y}_{\mathrm{HH-NA}}$. However, we will show in the next two examples that our conclusions may not always hold under larger departures from uniform response.

Suppose that there are exactly 1,188 mail-out non-respondents and that the values of C, $c^{(1)}$, $c^{(2)}$, $c^{(3)}$, $P_2^{(1)}$, $P_2^{(2)}$ and $P_2^{(3)}$ are exactly the same as those used in the simulation study and Table 4.1. However, for one of the 1,188 units, unit j say, the probabilities $P_{2j}^{(1)} = 0.25$, $P_{2j}^{(2)} = 0.05$ and $P_{2j}^{(3)} = 0.70$ are replaced with $P_{2j}^{(1)} = 0.000005$, $P_{2j}^{(2)} = 0.000001$ and $P_{2j}^{(3)} = 0.999994$, respectively. The response mechanism is almost uniform, except for one unit with a very small probability of being resolved. For simplicity, we assume that the follow-up sample is selected using simple random sampling without replacement. For this scenario, Table 4.2 shows the sample size (3.3), the expected number of respondents (3.4), the expected response rate (3.5) and the expected number of resolved units (3.6) for different values of K.

Table 4.1
Sample size, expected response rate, and expected number of respondents and resolved units for different values of K under uniform response to the follow-up

K	Sample size (3.7)	Expected response rate (3.9)	Expected number of respondents (3.8)	Expected number of resolved units (3.10)
∞	439	83.3%	366	439
20	439	83.3%	366	439
10	452	81.0%	366	439
6	498	73.5%	366	439
5	528	69.3%	366	439
4	578	63.3%	366	439
3	668	54.8%	366	439
2	861	42.5%	366	439
1*	1,188	25.0%	297	356

*The direct application of (3.7) leads to $n_2(C,1) =$ 1,463. However, this value is larger than the expected number of mail-out non-respondents in the simulation study. Assuming there are exactly 1,188 mail-out non-respondents and taking them all in the follow-up sample $(n_2 = 1{,}188)$, we can compute the expected follow-up cost (3.1) as $\tilde{C}(n_2,1) =$ 2,435.4, which is smaller than the total budget of 3,000. Using a revised budget of 2,435.4, the expected number of respondents and resolved units are 297 and 356, respectively.

Table 4.2
Sample size, expected response rate, and expected number of respondents and resolved units for different values of K when one unit has a very small probability of being resolved

K	Sample size (3.3)	Expected response rate (3.5)	Expected number of respondents (3.4)	Expected number of resolved units (3.6)
∞	20	83.3%	17	20
20	439	83.2%	365	438
10	452	80.9%	365	438
6	498	73.5%	366	439
5	528	69.3%	366	439
4	578	63.3%	366	439
3	668	54.7%	366	439
2	861	42.5%	366	439
1*	1,188	25.0%	297	356

*The direct application of (3.3) leads to $n_2(C,1) =$ 1,464, which is larger than the number of mail-out non-respondents (1,188). Similar to Table 4.1, we can compute the expected follow-up cost (3.1) as $\tilde{C}(n_2 = 1{,}188,\ K = 1) =$ 2,434.4, which is smaller than the total budget of 3,000. Using a revised budget of 2,434.4, the expected number of respondents and resolved units are 297 and 356, respectively.

The minimum sample size to expend the budget, on average, is $n_2(C,\infty)=20$ in that scenario. It is significantly smaller than 439, the corresponding value for uniform response shown in Table 4.1. As pointed out in Section 3, using a finite value of K may avoid spending too large a portion of the budget on a few units with a very small probability of being resolved (unit j in this example). Indeed, Table 4.2 shows that the expected response rate decreases marginally by reducing the value of K from infinity to 20 whereas the expected number of respondents drastically increases from 17 to 365. Using a finite value of K seems desirable in this scenario as it may substantially reduce the variance of $\hat{Y}_{\mathrm{HH-NA}}$. The impact on non-response bias is likely to be negligible unless the y value of unit j is extremely different from other units. Incorrectly assuming uniform response for all units would lead to choosing a sample size of 439, as shown in Table 4.1. This choice appears to remain appropriate for this non-uniform follow-up response mechanism.

Suppose again that there are 1,188 mail-out non-respondents, the values of C, $c^{(1)}$, $c^{(2)}$ and $c^{(3)}$ are the same as those used in the simulation study and Table 4.1, and the follow-up sample is selected using simple random sampling without replacement. Suppose now the 1,188 mail-out non-respondents can be divided into two response homogeneous groups, each of size 594. The probabilities are $P_{2hi}^{(1)}=0.45$, $P_{2hi}^{(2)}=0.09$ and $P_{2hi}^{(3)}=0.46$ for the 594 units in the first group and $P_{2hi}^{(1)}=0.05$, $P_{2hi}^{(2)}=0.01$ and $P_{2hi}^{(3)}=0.94$ for the remaining 594 units. The response mechanism is not uniform; it is uniform within each of the two response homogeneous groups. The average probabilities over the 1,188 mail-out non-respondents are the same as those given in the uniform response scenario. Table 4.3 shows the sample size (3.3), the expected number of respondents (3.4), the expected response rate (3.5), and the expected number of resolved units (3.6) for different values of K.

Table 4.3
Sample size, expected response rate, and expected number of respondents and resolved units for different values of K under uniform response within groups

K	Sample size (3.3)	Expected response rate (3.5)	Expected number of respondents (3.4)	Expected number of resolved units (3.6)
∞	235	83.3%	196	235
20	305	71.2%	217	261
10	409	60.9%	249	299
6	519	54.2%	281	338
5	566	51.9%	294	352
4	629	48.9%	308	370
3	727	44.7%	325	390
2	914	37.7%	344	413
1*	1,188	25.0%	297	356

*The direct application of (3.3) leads to $n_2(C,1)=$ 1,463, which is larger than the number of mail-out non-respondents (1,188). Similar to Table 4.1, we can compute the expected follow-up cost (3.1) as $\tilde{C}(n_2=$ 1,188, $K=1)=$ 2,435.4, which is smaller than the total budget of 3,000. Using a revised budget of 2,435.4, the expected number of respondents and resolved units are 297 and 356, respectively.

The minimum sample size to expend the budget, on average, is $n_2(C,\infty)=235$, which is much smaller than the corresponding value of 439 for uniform response. In this scenario, using a finite value of

K does not seem advantageous. By decreasing the value of K from infinity to 20, the expected number of respondents only increases by 21 whereas the expected response rate decreases by more than 10%. The small variance reduction could possibly be offset by a larger increase of non-response bias. The magnitude of non-response bias depends on the strength of the association between the y variable and the response homogeneous groups. A small value of K (a large sample size) might be appropriate if this association is weak so as to benefit from a larger expected number of respondents. However, this is a risky choice as the expected response rate would drop significantly, thereby offering a reduced protection against departure from the assumed response mechanism. Therefore, a sample size of 439 in this scenario might not be appropriate due to the increased risk of non-response bias. Then non-response bias can be dampened at the estimation stage, at least asymptotically, by computing the non-response weight adjustment (2.5) separately for each response homogeneous group. This weighting strategy is standard and should be used when response homogeneous groups can be identified; yet it does not offer full protection against departure from the assumed response mechanism. It is for this reason that a large value of K, even infinite, may be preferable in this scenario.

As pointed out in Section 3, plots of the expected response rate and the expected number of respondents as a function of K may be useful to determine a suitable trade-off between the maximization of the expected response rate $(K=\infty)$ and the maximization of the expected number of respondents, as illustrated in the above examples. An infinite value of K should be the default as it minimizes non-response bias. However, a large finite value of K might be appropriate if it sharply increases the expected number of respondents with minimal impact on the expected response rate.

5. Conclusions

In Section 3, we derived an explicit expression for $n_2(C,\infty)$, the minimum sample size to expend the budget C, on average, while resolving all follow-up sample units. We showed that this minimum sample size maximizes the expected response rate; thereby minimizing the bias of the non-response-adjusted Hansen and Hurwitz (1946) estimator. Our empirical investigations showed that this minimum sample size also appears to minimize the mean square error of this estimator. This can be explained by noting that the expected number of respondents remain roughly constant as the sample size increases, yielding an approximately constant variance. For the uniform follow-up response mechanism, it was possible to show theoretically that the expected number of respondents does not vary as the sample size increases (or does not vary with K), confirming the empirical results.

At first glance, the idea of maximizing the expected response rate to minimize non-response bias may appear to contradict existing non-response literature. It is well known that a data collection procedure that intends to maximize the response rate for a given sample will most likely increase the non-response bias when easy-to-reach respondents differ from the other sample units. That is, increasing the response rate does not necessarily reduce non-response bias for a given sample and may actually do the opposite. Our

results do not contradict this statement as we studied a different feature of the data collection design: the effect of the follow-up sample size on the expected response rate and non-response bias. It appears that this question has not been investigated in the literature. Our main conclusion is that a smaller follow-up sample size contributes to increasing the expected response rate and decreasing non-response bias.

Our conclusions may have important implications in practice. In business surveys conducted by Statistics Canada, all the mail-out non-respondents are currently followed up, and an adaptive collection procedure is used to prioritize cases (see Bosa et al., 2018). We believe that the non-response bias could be further reduced by following up only a sample of mail-out non-respondents in situations where the follow-up budget is insufficient to properly handle the volume of mail-out non-respondents. The adaptive collection procedure currently in place could continue to be used to manage data collection of the follow-up sample.

Another conclusion of our empirical investigations is that the PPS designs appeared to perform slightly better than the SRS and stratified SRS designs. However, no attempt was made to optimize the stratification or allocation of the stratified SRS design. The performance of the stratified design would likely be improved through a more efficient use of the auxiliary variable “Revenue” for stratification.

Finally, we observed that, unlike the follow-up response mechanism, the mail-out response mechanism had no impact on the bias of the non-response-adjusted Hansen and Hurwitz (1946) estimator. As a result, the mail-out non-response bias could be eliminated, even if the mail-out response probability was correlated to the variable of interest, provided that the follow-up response probability was uniform. This result is not surprising since the estimator of Hansen and Hurwitz (1946) is unbiased for any mail-out response mechanism.

Acknowledgements

The authors would like to thank three anonymous referees and the Associate Editor for their constructive comments, which led to significant improvements to the clarity of the manuscript.

References

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

Beaumont, J.-F., 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, Statistics Canada, May 2014, Ottawa.

Bosa, K., Godbout, S., Mills, F. and Picard, F. (2018). How to decompose the non-response variance: A total survey error approach. *Survey Methodology*, 44, 2, 291-308. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2018002/article/54957-eng.pdf.

Groves, R.M., and Heeringa, S.G. (2006). Responsive design for household surveys: Tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society, Series A*, 169, 439-457.

Hansen, M.H., and Hurwitz, W.N. (1946). The problem of non-response in sample surveys. *Journal of the American Statistical Association*, 41, 517-529.

Peytchev, A., Riley, S., Rosen, J., Murphy, J. and Lindblad, M. (2010). Reduction of non-response bias in surveys through case prioritization. *Survey Research Methods*, 4, 21-29.

Schouten, B., Cobben, F. and Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, 35, 1, 101-113. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2009001/article/10887-eng.pdf.

Statistics Canada (2017). *Monthly Survey of Food Services and Drinking Places (MSFSDP)*. Statistics Canada, http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=413027.

Thompson, K.J., Kaputa, S. and Bechtel, L. (2018). Strategies for subsampling nonrespondents for economic programs. *Survey Methodology*, 44, 1, 75-99. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2018001/article/54929-eng.pdf.

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, Series A*, 180, 203-223.

Xie, H., Godbout, S., Youn, S. and Lavallée, P. (2011). Collection Follow-Up Operation Using Priority Scores For Business Surveys. Conference of European Statisticians, Work Session on Statistical Data Editing, Ljubljana (Slovenia).

Survey Methodology, June 2022
Vol. 48, No. 1, pp. 119-144
Statistics Canada, Catalogue No. 12-001-X

Using Multiple Imputation of Latent Classes to construct population census tables with data from multiple sources

Laura Boeschoten, Sander Scholtus, Jacco Daalmans, Jeroen K. Vermunt and Ton de Waal[1]

Abstract

The Multiple Imputation of Latent Classes (MILC) method combines multiple imputation and latent class analysis to correct for misclassification in combined datasets. Furthermore, MILC generates a multiply imputed dataset which can be used to estimate different statistics in a straightforward manner, ensuring that uncertainty due to misclassification is incorporated when estimating the total variance. In this paper, it is investigated how the MILC method can be adjusted to be applied for census purposes. More specifically, it is investigated how the MILC method deals with a finite and complete population register, how the MILC method can simultaneously correct misclassification in multiple latent variables and how multiple edit restrictions can be incorporated. A simulation study shows that the MILC method is in general able to reproduce cell frequencies in both low- and high-dimensional tables with low amounts of bias. In addition, variance can also be estimated appropriately, although variance is overestimated when cell frequencies are small.

Key Words: Combined survey-register data; Population census; Misclassification; Multiple imputation; Latent Class analysis.

1. Introduction

Official Statistics are increasingly often compiled from a combination of data sources, including surveys and administrative registers. The use of different sources poses multiple challenges. Different sources can be overlapping, meaning that more than one observation is obtained for the same person and variable. Often, it is observed that data sources are contaminated by errors and missing values. Therefore it can happen that two data sources provide two different values for the same unit and variable. Most of the data collected by statistical agencies have to be corrected or processed somehow to obtain consistent and publishable results. Several strategies are available to deal with multiple, overlapping data sources that are each contaminated by erroneous and missing values, see e.g. Pankowska, Pavlopoulos, Bakker and Oberski (2020). A first, and in practice often chosen strategy, is to ignore inconsistencies between data sources. This happens for instance if one data source is chosen that is believed to have the highest quality (de Waal, van Delden and Scholtus, 2020). When such strategies are chosen, the information in all available sources is not fully exploited.

A second strategy is to apply weighting techniques (Särndal, Swensson and Wretman, 2003). When weighting is used, survey records are calibrated towards the totals from a register source. Differences between data sources are fully explained from the selection effects of the sample. This approach ignores

1. Laura Boeschoten, Tilburg University Tilburg School of Social and Behavioral Sciences - Methodology and Statistics Warandelaan 2 Tilburg, Tilburg, Noord-Brabant 5000 LE Netherlands and Centraal Bureau voor de Statistiek - Methodology Henri Faasdreef 312, Den Haag 2490 HA Netherlands. E-mail: lauraboeschoten@gmail.com; Sander Scholtus, Centraal Bureau voor de Statistiek - Methodology Den Haag, Zuid-Holland Netherlands; Jacco Daalmans, Centraal Bureau voor de Statistiek - Methodology Den Haag, Zuid-Holland Netherlands; Jeroen K. Vermunt, Tilburg University Tilburg School of Social and Behavioral Sciences - Methodology & Statistics Tilburg, Noord-Brabant Netherlands; Ton De Waal, Centraal Bureau voor de Statistiek - Methodology Den Haag, Zuid-Holland Netherlands and Tilburg University Tilburg School of Social and Behavioral Sciences - Methodology & Statistics Tilburg, Noord-Brabant Netherlands.

the fact that the register totals, as well as the sample surveys, might be subject to measurement error. An additional complication is that weighting does not always lead to fully consistent output, as it only achieves consistency with regard to the variables that are incorporated in the weighting model. The number of variables that can be included in a weighting model is however limited.

A third strategy to resolve inconsistencies between multiple sources is macro-integration, an approach that reconciles statistical output at aggregate level. This approach usually consists of two steps. First, differences with a known cause are resolved (i.e. bias). The remaining, mostly smaller, discrepancies that usually arise due to noise are corrected in a second step. Several mathematical methods have been developed for this purpose, e.g. Bikker, Daalmans and Mushkudiani (2013), Daalmans (2019), Di Fonzo and Martini (2003), Magnus, van Tongeren and de Vos (2000), Sefton and Weale (1995) and Stone, Champernowne and Meade (1942). A first drawback of macro integration is that the connection between the micro-data and the published results gets lost. The macro-integrated results cannot be computed by aggregation of the micro data. A second drawback is that the detailed micro data might not be fully exploited, as the corrections are made at the macro level.

Many of the issues arising when one of the previously discussed strategies is used can be circumvented by Multiple Imputation of Latent Class analysis (MILC) by Boeschoten, Oberski and de Waal (2017). This method combines multiple measures from different sources (population register and sample survey) at micro level. The different observations are considered indicators of a Latent Class (LC) model. The MILC-model corrects for misclassification while also taking edit restrictions into account. These are rules that identify logically impossible combinations of scores (e.g. pregnant men). After the LC model has been estimated, multiple imputed versions of the target variable are created, that are corrected for the estimated misclassification. Differences between imputed values reflect the uncertainty due to missing and conflicting values. The total variance can be estimated based on these differences. The method can be considered a model-based imputation method that requires the Missing At Random (MAR) assumption. A simulation study on the performance of this method showed that its performance is strongly related to the entropy R^2 value of the LC model; a measure which indicates how well the LC model can predict class membership based on the observed variables, or how well classes are separated.

After MILC was introduced, multiple studies have extended the method to broaden its scope of applicability. Boeschoten, de Waal and Vermunt (2019) extended the method to impute values that are missing by design, for example because they were not present in the sample, using a quasi-latent variable. More specifically, a quasi-latent variable is a latent variable that is restricted to have a perfect relationship with an observed variable that contains missing values. In that way, the relationship between the quasi-latent variable and all other variables specified in the model can be used to estimate the missing values. In addition, they investigated the performance of the method when two combined sources follow different missingness mechanisms. Furthermore, Boeschoten, Filipponi and Varriale (2021) investigated how the method can be extended for longitudinal situations and how unit missingness can be imputed in a situation of combined survey and register data.

Although these previous studies investigated a number of relevant issues, there are still cases for which it is unclear how the MILC-method can be applied. The aim of this paper is to further enhance the possibilities of MILC in terms of application and, with that, to further increase the capabilities of producing multi-source statistics.

Currently, the application of MILC has been limited to univariate problems. In practice, however, there is often a need to estimate multiple variables at once. The first important extension in this paper is to allow the simultaneous imputation of multiple latent variables. As population registers can contain misclassification, it is worthwhile to correct for the misclassification if possible. For multivariate problems, corrections should be performed simultaneously, which is more difficult than for one variable only.

Second, statistical agencies generally consider finite target populations (e.g. containing all registered inhabitants of a country). It is unclear if the MILC method can be applied directly to a finite population, or that adaptations to the method should be made.

The usefulness of the extensions in this paper is illustrated by an application to the Dutch virtual census; an application that would otherwise not be possible. For the census, a large number of tables have to be estimated from a population register and a sample survey. To the best of our knowledge, this is the first time that MILC has been applied to such a large estimation problem. Theoretically, it is already known that edit restrictions can be incorporated in an LC model to prevent the occurrence of logically impossible combinations of scores (Boeschoten et al., 2017). However, it is not trivial how the MILC method performs if edit restrictions are incorporated in such a way that they affect multiple cells in a population census table.

In Section 2, a description of the MILC method is given, tailored to handle the specific extensions discussed. In Section 3, a description of the simulation study is given. Simulation results are shown in Section 4 and Section 5 provides a discussion.

2. Methodology

When applying the MILC method, the starting point is a unit-linked combined dataset, which can consists of combinations of administrative population registries and survey samples. In order to account for uncertainty regarding the parameters of the LC model estimated at a later step in MILC, a non-parametric bootstrap procedure is applied on this dataset first (step 1). This involves creating M bootstrap samples by drawing observations from the observed dataset with replacement. Subsequently, for each bootstrap sample, the LC model of interest is estimated (step 2) using Latent GOLD software (Vermunt and Magidson, 2013a). Here, model parameters are estimated by Maximum Likelihood using a combination of the Expectation-Maximization and Newton-Raphson algorithms. Note that here, by explicitly stating which cells should be restricted, constrained estimation is used. Next, M imputations are created using the M sets of parameter values obtained from the M latent class models (step 3). If imputations would be created based on the maximum-likelihood estimates obtained directly using the

original observed data, sampling uncertainty regarding the estimated parameters of the latent class model would be ignored.

In the following subsections, we explain each of the steps of MILC in more detail and present the extension for the estimation of multiple latent variables for a finite population from register and sample survey data.

2.1 Step 1: Creating bootstrap samples

We propose to use the "classical" bootstrap procedure here, which consists of repeatedly drawing samples with replacement from the original dataset, of the same size as the original dataset. A motivation for using this classical with-replacement bootstrap here, as opposed to an adapted bootstrap procedure for a finite population, is provided in Section 2.5 below.

The bootstrap should be applied to the dataset that is used to estimate the LC models. When register data and survey data are combined, the indicator variables from the survey will typically be missing for a large part (e.g., 90% or more) of the population. The LC models could then be estimated by two different approaches:

- using only the subset of persons observed in both the survey and the register (complete cases);
- using all available data, including cases with missing indicators.

Under the second approach, full information maximum likelihood can be used to handle missing values when estimating the LC models. This has the advantage of using all available information. Since this amounts to estimating the LC model on M datasets with the size of the target population, a practical drawback of this approach is that it may be computationally demanding in terms of time and memory. Therefore, the first approach may be more attractive, in particular when the associations among the covariates and target variables are relatively weak. In the latter approach, the cases with missing survey data will contain relatively little information about the parameters of the LC model. Note that under both approaches, the estimated LC models are used to impute predictions of the latent classes throughout the population. Depending on which approach is chosen to estimate the LC models, bootstrapping is applied either to the subset of complete cases or to the target population. In the simulation study in this paper, the complete-case approach will be used.

2.2 Step 2: Estimating the latent class model

The second step performed is the estimation of the LC model. It is explained below how this is done for multiple latent variables. As described in the previous section, the LC model is typically estimated M times using the M bootstrapped datasets. In the situation under evaluation in this paper, the LC model is estimated M times on M subsets of complete observations coming from the M bootstrap samples. An extensive discussion of the model and the assumptions made when using the model to correct for measurement error can be found in Boeschoten et al. (2017). Multiple latent variables can be estimated

simultaneously in one model, which yields the following model structure for the joint probability of the response variables given covariate values, denoted by $\mathrm{P}\left(\mathbf{Y}=\mathbf{y} \mid \mathbf{Q}=\mathbf{q}\right)$. The number of latent variables is denoted as v and K_h is the number of classes of latent variable X_h (scalar), where $\left(h=1,\ldots,v\right)$. Furthermore, $\mathbf{Y}$ are the observed target variables, i.e. the indicator variables, L_h is the number of indicator variables for X_h and $\mathbf{Q}$ are the (also observed) covariate variables:

$$\mathrm{P}\left(\mathbf{Y}=\mathbf{y} \mid \mathbf{Q}=\mathbf{q}\right) = \sum_{x_1=1}^{K_1} \ldots \sum_{x_v=1}^{K_v} \mathrm{P}\left(X_1=x_1,\ldots,X_v=x_v \mid \mathbf{Q}=\mathbf{q}\right) \prod_{l_1=1}^{L_1} \mathrm{P}\left(Y_{l_1,1}=y_{l_1,1} \mid X_1=x_1\right) \ldots \prod_{l_v=1}^{L_v} \mathrm{P}\left(Y_{l_v,v}=y_{l_v,v} \mid X_v=x_v\right). \tag{2.1}$$

Here, local independence is assumed as well as independence of covariates.

Constrained parameter estimation is used when certain cells within $\mathrm{P}\left(X_1=x_1,\ldots,X_v=x_v \mid \mathbf{Q}=\mathbf{q}\right)$ are restricted. This can be used to specify that certain combinations of scores between covariates and latent variables are logically impossible, or when a "quasi-latent" variable is used to create imputations for missing values in a variable (Vermunt and Magidson, 2013b).

2.3 Step 3: Multiple imputation

To be able to create multiple imputations, joint posterior membership probabilities are calculated for every person in the original dataset. They represent the probability that a unit is part of a combination of latent classes from the different latent variables, given its combination of scores on the indicators and covariates used in the LC model. These probabilities can be used to create multiple imputations of the latent variables which contain their "true scores".

The joint posterior membership probabilities can be calculated by applying Bayes' rule to the conditional response probabilities obtained from the M LC models:

$$\mathrm{P}\left(X_1=x_1,\ldots,X_v=x_v \mid \mathbf{Y}=\mathbf{y},\mathbf{Q}=\mathbf{q}\right) = \frac{\mathrm{P}\left(X_1=x_1,\ldots,X_v=x_v,\mathbf{Y}=\mathbf{y} \mid \mathbf{Q}=\mathbf{q}\right)}{\mathrm{P}\left(\mathbf{Y}=\mathbf{y} \mid \mathbf{Q}=\mathbf{q}\right)}, \tag{2.2}$$

where

$$\mathrm{P}\left(X_1=x_1,\ldots,X_v=x_v,\mathbf{Y}=\mathbf{y} \mid \mathbf{Q}=\mathbf{q}\right) = \mathrm{P}\left(X_1=x_1,\ldots,X_v=x_v \mid \mathbf{Q}=\mathbf{q}\right) \prod_{l_1=1}^{L_1} \mathrm{P}\left(Y_{l_1,1}=y_{l_1,1} \mid X_1=x_1\right) \ldots \prod_{l_v=1}^{L_v} \mathrm{P}\left(Y_{l_v,v}=y_{l_v,v} \mid X_v=x_v\right) \tag{2.3}$$

and $\mathrm{P}\left(\mathbf{Y}=\mathbf{y} \mid \mathbf{Q}=\mathbf{q}\right)$ is defined in equation 2.1. For one profile (so one set of scores on all indicator and covariate variables), the joint posterior membership probabilities sum up to one.

To be able to include parameter uncertainty in our variance estimates, we perform the model estimation on M bootstrap samples of the dataset, resulting in M different LC models. We generate imputations in the original dataset accounting for the parameter uncertainty by using the resulting M sets of bootstrap parameter estimates. More specifically, with each of these M parameter sets we compute the posterior class membership probabilities for the original sample, and use these to generate the imputations. In other words, the M imputations are based of M different sets of posterior probabilities.

2.4 Step 4: Pooling

The next step is to obtain estimates of interest for every imputation, and to pool them using Rubin's Rules (Rubin, 1987, page 76). For this research, the main interest is producing a frequency table. Therefore, the frequency table of interest is obtained for the M imputations and they are pooled, which means taking the average over the imputations for every cell in the frequency table:

$$\hat{\theta}_j = \frac{1}{M}\sum_{i=1}^{M}\hat{\theta}_{ij}, \tag{2.4}$$

where j refers to a specific cell in the frequency table.

Next, an estimate of the uncertainty around these frequencies is of interest. In general, the variance of the pooled estimate j can be estimated by Rubin's total variance formula for multiple imputation (Rubin, 1987, page 76):

$$\mathrm{VAR}_{\mathrm{total}_j} = \overline{\mathrm{VAR}}_{\mathrm{within}_j} + \mathrm{VAR}_{\mathrm{between}_j} + \frac{\mathrm{VAR}_{\mathrm{between}_j}}{M}. \tag{2.5}$$

Here, $\mathrm{VAR}_{\mathrm{between}_j}$ can be estimated as

$$\mathrm{VAR}_{\mathrm{between}_j} = \frac{1}{M-1}\sum_{i=1}^{M}\left(\hat{\theta}_{ij}-\hat{\theta}_j\right)\left(\hat{\theta}_{ij}-\hat{\theta}_j\right)'. \tag{2.6}$$

The within variance $\mathrm{VAR}_{\mathrm{within}_j}$ reflects the average sampling variance of ij when the imputed values are treated as observed. In our application, as the population is finite and imputations are generated for the complete population, this within variance component is zero and can be mitigated (Vink and van Buuren, 2014). Note that this is a property of multiple imputation and is due to the fact that the complete population is imputed. This should not be confused with the decision to only use a sample for LC model estimation. Hence, formula (2.5) is reduced in this case to:

$$\mathrm{VAR}_{\mathrm{total}_j} = \mathrm{VAR}_{\mathrm{between}_j} + \frac{\mathrm{VAR}_{\mathrm{between}_j}}{M}. \tag{2.7}$$

2.5 A note on bootstrapping for multiple imputation in finite populations

The aim of a census is to estimate certain target parameters of a finite population (e.g., all persons currently living in the Netherlands). Hence, a natural idea might be to apply a finite-population bootstrap

procedure in this context; see Mashreghi, Haziza and Léger (2016) for an overview of bootstrap methods for finite populations. However, when determining the appropriate bootstrap approach, it should be noted that the bootstrap in MILC is specifically implemented to account for the between imputation variance component of formula (2.5) in Section 2.4. In general, variability in the target parameters due to the fact that a sample was drawn from a finite population is incorporated in the within variance component of formula (2.5). As we use mass imputation here, the within variance component in fact reduces to zero; cf. formula (2.7). More generally, this component would be estimated separately from the bootstrap method at hand; see Boeschoten et al. (2017) for an example.

Furthermore, the reason for incorporating the bootstrap in the MILC approach is to account for uncertainty in the estimated parameters of the latent class model. Note that these parameters are not associated with a finite population, but with a model. Even if we had observed the entire finite population, there would still be uncertainty about the true parameter values of the latent class model. This uncertainty can be considered as drawing from an infinite distribution. Therefore, we select the classical with-replacement bootstrap. We argue that bootstrap methods for finite populations should not be used in this context. For large samples, such methods would result in a substantial underestimation of the variance when combined with the usual approach to multiple imputation. We also checked this empirically in the simulation study to be discussed in Section 3. As an example, when a pseudo-population bootstrap method for finite populations was used, the resulting se/sd ratios in Table 4.7 for the condition MAR, $M = 5$ were 0.7217, 0.7887, 0.7536 and 0.8607, respectively, all pointing to a non-negligible underestimation of the true variance.

In the simulation study in this paper, we will restrict attention to surveys based on simple random sampling and stratified simple random sampling. For more complex survey designs, e.g. involving cluster sampling or sampling with unequal probabilities, it is unclear whether the proposed bootstrap approach is always appropriate. It is possible that in some cases such complex design features could indirectly affect the uncertainty of estimated parameters of the latent class model and therefore become relevant for variance estimation. We will return to this point in the discussion section.

3. Simulation study

In this section, we describe a simulation study that is performed to evaluate the extensions of the MILC method in Section 2. The topic of this study is the estimation of a table from the Dutch Population and Housing Census.

3.1 The Dutch Census

Population and housing censuses provide a picture about the socio-demographic and socio-economic situation of a country and it is ubiquitous that a census should cover the entire population of people and dwellings that are present in a country. Every ten years the United Nations Economic and Social Council (ECOSOC) adopts a resolution, urging Member States to carry out a population and housing census and to

disseminate census results as an essential source of information, see e.g. The Economic and Social Council (2005). In the EU, explicit agreements have been made about which variables should be listed in the census, and also which cross-tables should be produced (European Commission, 2008, 2009 and 2010).

The vast majority of countries produce census data by conducting a traditional census, which entails interviewing inhabitants in a complete enumeration, reaching every single household. An increasing number of countries however have adopted a different, innovative approach, in the form of a so-called virtual census. With a virtual census, census tables are compiled using data sources that are already available at the statistical agency. These are data sources that have not been primary collected for the census, but for other purposes. Statistics Netherlands can rely on population registers as the main source for most census tables. These registers are of relatively good quality, including a very broad coverage (Geerdinck, Goedhuys-van der Linden, Hoogbruin, De Rijk, Sluiter and Verkleij, 2014). All register variables are available from Statistics Netherlands' system of social statistical data-sets (Bakker, Van Rooijen and Van Toor, 2014). The backbone is the Central Population Register which combines the population registers from municipalities. The population registers are supplemented with variables originating from sample surveys, because not all variables that are necessary according to the EU regulations can be found in the population registers.

For the 2001 and 2011 Dutch censuses, only two variables could not be measured from registers: Occupation and Educational Attainment (Schulte Nordholt, Van Zeijl and Hoeksma, 2014). These two variables were observed from combined Labour Force Surveys (LFSs). To obtain the required cross-tables for the 2011 Dutch census, a procedure was used where all data sources were matched on the unit level. Then, a micro-integration process was carried out. Micro-integration brings together records from different micro-datasets and subsequently resolves data inconsistencies. The goal is to improve the quality, compatibility and scope of the data sets. The techniques that are used in micro integration are: completing, harmonising and correcting for measurement errors. Completing means that corrections are made for an under- or overcoverage of a target population. Harmonisation refers to transformations such that data sets fit to the concept that is supposed to be measured. Measurement correction means that inconsistencies between sources are resolved (Bakker, 2011; van Rooijen, Bloemendal and Krol, 2016). Also, inconsistencies between sources are removed, by using formal rules that make clear what happens in case of inconsistencies, e.g. which source is used (Bakker, 2010; de Waal, Pannekoek and Scholtus, 2011).

After micro-integration, two combined data sources were obtained: one based on a combination of registers and the other one based on a combination of sample surveys. All census tables that do not contain occupation and educational attainment were entirely compiled from the combined registers. The values in the cells of these tables were obtained by counting the occurrence of the categories in the matched registers. The other census tables, those with educational attainment and/or occupation, were estimated from the combined sample surveys. To establish consistent results, a procedure was applied based on weighting followed by macro integration (Daalmans, 2018; Schulte Nordholt et al., 2014). In the first step,

weights were derived, such that the marginal totals of the weighted survey data comply with the known totals from the registers. The different tables that are obtained in this way are not necessarily consistent with each other, because different weighting schemes apply to each table. To resolve this problem, macro-integration is used. This step starts with initial estimates for each census table, derived from the weighted survey data or from the integrally counted register data. These initial estimates are adjusted, to arrive at fully consistent census tables, that comply with the known register totals.

MILC has a couple of advantages over the current estimation method. First, the assumption is often made that the population registers are free of error. If a variable is measured both in the population register and in a sample survey and the scores on these variables contradict each other, the register score usually overrides the survey score because of this assumption. In other words, sample survey data are ignored for the part that is also observed in a register. Second, for the current procedure, it is not easy to compute uncertainty measures that capture all steps of the estimation process, including the uncertainty due to the missing and conflicting values in the linked data-sets. For MILC on the other hand it is well-established how variances can be properly estimated. Third, the data processing procedure that is currently used contains a specific sequence of steps, where decisions made at one step are influenced by decisions made at previous steps. For instance, if there are two conflicting values for the same person, then one of these is chosen in the "micro-integration" step. In the subsequent weighting and macro integration steps only one value is used. Thus, the availability of the different values is ignored in the final estimation of the census tables. Basically, MILC exploits information provided by all observed values in contrast to the current procedure.

3.2 The census table under investigation

The starting point of this simulation study is an existing census table, which can be downloaded from Census Hub (Census Hub, 2017). This table comprises 2,691,477 persons who where living in the region "Noord-Holland" in the Netherlands in 2011. This census table is a cross-table between the following six variables:

1. Age in 21 categories: under 5 years; 5 to 9 years; 10 to 14 years; 15 to 19 years; 20 to 24 years; 25 to 29 years; 30 to 34 years; 35 to 39 years; 40 to 44 years; 45 to 49 years; 50 to 54 years; 55 to 59 years; 60 to 64 years; 65 to 69 years; 70 to 74 years; 75 to 79 years; 80 to 84 years; 85 to 89 years; 90 to 94 years; 95 to 99 years; 100 years and over.
2. Marital status in eight categories: never married; married; widowed; divorced; registered partnership; widow of registered partner; divorced from registered partner; not stated.
3. Gender in two categories: male; female.
4. Place of birth in five categories: the Netherlands; a country within the European Union; a country outside the European Union; other; not stated.
5. Type of family nucleus in which a person lives in five categories: partners; lone parents; sons/daughters; not stated; not applicable.

6. Country of citizenship in five categories: Dutch citizen; citizen of a country within the European Union; citizen of a country outside the European Union; stateless; not stated.

Thus, the census table consists of 42,000 cells.

3.3 Simulation setup

The goal of this simulation study is to replicate the frequencies of the 42,000 cells in the cross-table using multiple indicators contaminated with misclassification and missing values. Therefore, this misclassification should be induced first.

We generate two indicator variables for three different latent variables, all containing 5% random misclassification, which can be considered a very high amount, especially for Dutch population registers. The indicator variables are generated for the variables "Gender", "Type of family nucleus" and "Country of citizenship". Misclassification is generated in such a way that first, 5% of the cases are randomly selected. Second, their original score is identified and third, a different score is assigned by sampling from the observed frequency distribution of the other categories.

For the register indicators $Y_{l_v,1}$, misclassification is generated only once, as these indicator variables represent register variables for the complete and finite population, there should not be any variability in misclassification between replications in the simulation study for these variables. For the survey indicators $Y_{l_v,2}$, misclassification is newly generated for every replication in the simulation study, followed by generating missing values using either a Missing Completely At Random (MCAR) or Missing At Random (MAR) missingness mechanism with approximately 90% missingness for both situations. With a MCAR mechanism, the response probabilities for the respondents and non-respondents is equal. With a MAR mechanism, the response probabilities are related to other observed values (Rubin, 1976). These $Y_{l_v,2}$ indicators represent survey variables for a sample of the population.

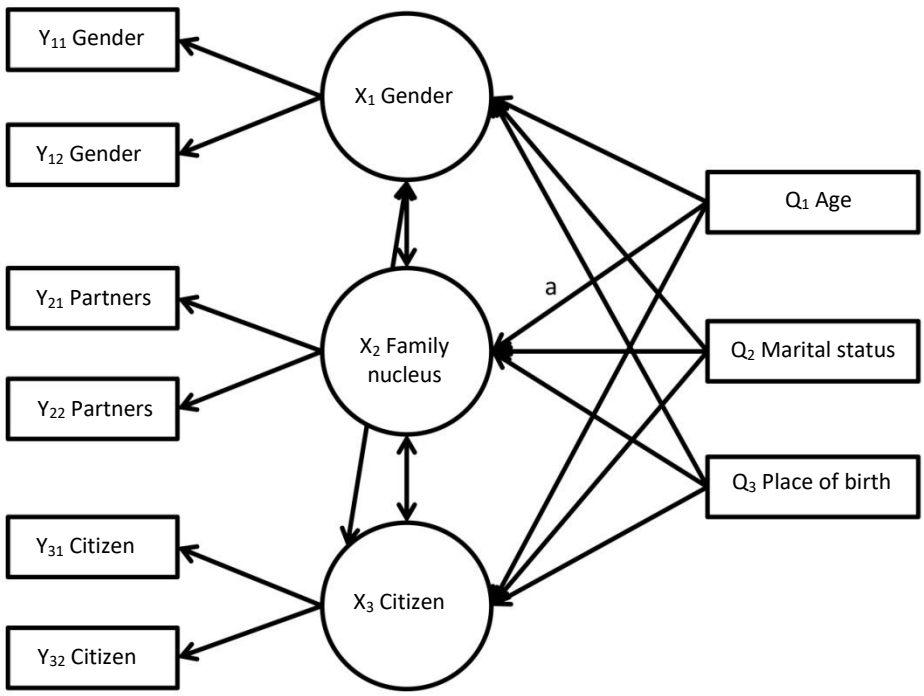
Missingness is generated in such a way that it mimics a situation that 10% of the population is included in the survey. Missingness is generated under MCAR and MAR. Under MCAR, the probability of being missing (i.e. not being included in the survey) is 90% and equal for every person in the population. Under MAR, the probability of being missing depends on a persons' age and decreases as a person gets older. More specifically, the probability of being missing is lowest for persons in the age category "100 years and older", and is 80%. This percentage gradually increases with the highest being 94% for the persons in the age category "under five years". To summarize, for each of the 500 iterations in the simulation study, a simple random or stratified sample of the combined data-set is obtained that contains approximately 269,147 persons (10% of the population), on which the LC model is estimated.

3.4 Applying the MILC method

As discussed in Section 2, M bootstrap samples are generated from the combined dataset, and in this study the LC model is estimated only on the complete set of observations of each bootstrap sample. Results are obtained using $M = 5$, $M = 10$ and $M = 20$.

In Figure 3.1, the graphical overview of the latent class model can be found. Here, it can be seen that the latent variables X_1 "Gender", X_2 "Family nucleus" and X_3 "Citizen" are all measured by two indicators. The restriction on the relationship between Q_1 "Age" and X_2 "Family nucleus" is denoted by "a" in Figure 3.1. Here, we restricted that if someone is of age category "under 5 years", "5 to 9 years" or "10 to 14 years", it is impossible to be assigned to the latent classes "partners" or "lone parents" for the latent variable "Family nucleus".

Figure 3.1 Graphical overview of the LC model specified. Note that edit restrictions are applied between the variables "Type of family nucleus" and "age" (denoted in the model by "a").



To specify the LC model for response pattern $\mathrm{P}\left(\mathbf{Y}=\mathbf{y} \mid \mathbf{Q}=\mathbf{q}\right)$ we can fill in at equation 2.1 that $v=3,\ K_1=2,\ K_2=4,\ K_3=4,\ L_1=2,\ L_2=2$ and $L_3=2.$ Note that X_2 here only has four latent classes, while the variable "Family Nucleus" in the population census table has five categories. Therefore, it would have made sense for X_2 to also have five latent classes. However, there were no observations for the category "not applicable", so therefore we didn't have to include a latent class for this category. The same holds for the category "stateless" of X_3.

Next, multiple imputations can be created and estimates of interest can be pooled as described in Sections 2.3 and 2.4. As the cells of the frequency-tables of interest can become very small, a log-transformation is used to ensure appropriate confidence intervals around these small cells. Therefore, $\mathrm{VAR}_{\mathrm{between}_j}$ is not estimated as the variance of $\hat{\theta}_{ij}$, as in equation 2.7, but as the variance of $\log\left(\hat{\theta}_{ij}\right)$, where $\hat{\theta}_{ij}$ refers to the number of units in cell j in imputation i.

3.5 Evaluation

To evaluate the performance of the MILC method when trying to construct the census table initially used to create the misclassified data, it is useful to make comparisons to results obtained when the variable observed in the register is used directly to create cross-tables. We refer to these results as obtained using $Y_{v,1}$. These results are equal over the 500 simulation iterations and the bias here directly reflects the misclassification in this indicator, which becomes more severe as the categories are more imbalanced in size due to the misclassification mechanism. Furthermore, it would be difficult to draw general conclusions from results obtained by only evaluating every single of the 42,000 cells of the complete census table. Therefore, we investigate some specific characteristics of this table separately. First, we investigate whether the method is able to reconstruct the univariate marginal cell frequencies of the latent variables specified. Second, we investigate if the method is able to reconstruct the joint distribution of the three latent variables. Third, we investigate if the method correctly incorporates edit restrictions. At last, we investigate some features of the complete census table.

First, we evaluate the cell-proportions of the previously discussed cross-tables in terms of bias, by evaluating the average absolute bias and the root mean squared error (RMSE) over the $N_{it} = 500$ replications in the simulation study. More specifically, the bias of a cell frequency θ_j is calculated as

$$\text{bias}_{\theta_j} = \frac{\sum_{it=1}^{N_{it}} \left(\theta_j - \hat{\theta}_{j_{it}} \right)}{N_{it}}. \tag{3.1}$$

Furthermore, the RMSE is calculated as

$$\text{RMSE} = \frac{\sqrt{\sum_{it=1}^{N_{it}} \left(\theta_j - \hat{\theta}_{j_{it}} \right)^2}}{N_{it}}. \tag{3.2}$$

Second, results are evaluated in terms of variance. Here, it is of interest to evaluate whether MILC correctly reflects uncertainty due to missing and conflicting values in between imputation variance for both univariate and multivariate cross-tables. Therefore, we investigate if the average of the estimated standard errors is approximately equal to the standard deviation over the 500 estimates obtained from the 500 simulation replications by evaluating its ratio, which is calculated by

$$\frac{\left[\sum_{it=1}^{N_{it}} \text{SE}\left(\hat{\theta}_{j_{it}} \right) \Big/ N_{it} \right]}{\text{SD}\left(\hat{\theta}_{j_{it}} \right)}, \tag{3.3}$$

where SE is the square root of the estimate of the total variance obtained after applying pooling rules (Rubin, 1976) and $\text{SD}\left(\hat{\theta}_{j_{it}} \right)$ is calculated as

$$\text{SD}\left(\hat{\theta}_{j_{it}} \right) = \sqrt{\frac{\sum_{it=1}^{N_{it}} \left(\hat{\theta}_{j_{it}} - \bar{\theta}_{j_{it}} \right)^2}{N_{it}}}. \tag{3.4}$$

To account for small cell frequencies, $\hat{\theta}_{j_{it}}$ and $\bar{\theta}_{j_{it}}$ are considered on a log scale in equations 3.2, 3.3 and 3.4. To summarize, we denote the specific conditions evaluated in this simulation study as $Y_{v,1}$, MILC-MCAR-5, MILC-MCAR-10, MILC-MCAR-20, MILC-MAR-5, MILC-MAR-10 and MILC-MAR-20.

4. Simulation results

First, cell-proportions of univariate and multivariate cross-tables are evaluated in terms of bias and root mean squared error (RMSE) over the 500 simulation replications. Second, these cell-proportions are evaluated in terms of variance by investigating the average of the estimated standard error divided by the standard deviation over the 500 estimates obtained from the 500 simulation replications (SESD). Due to the log-transformations we made in equations 3.2, 3.3 and 3.4 to account for small cell frequencies, the RMSE and SESD are reported on a log scale.

4.1 Results in terms of bias

4.1.1 Univariate marginal frequencies of imputed variables

In Table 4.1, the simulation results can be found that cover the univariate marginal frequencies of the imputed latent variable "Gender" in terms of bias and RMSE. Results from all simulation conditions are shown. Here, it can be seen that a smaller amount of bias is obtained if $Y_{1,1}$ is used, compared to results obtained using MILC under all conditions. In addition, it can be seen that the RMSE is also smaller if $Y_{1,1}$ is used instead of the MILC method. Furthermore, it can be seen that both bias and RMSE slightly decrease as M increases, and that the quality of the results appears to be unrelated to the missingness mechanism.

Table 4.1
Results in terms of bias and root mean squared error for the two categories of the imputed latent variable "Gender"

				MCAR			MAR		
	Gender	**Frequency**	$Y_{1,1}$	$M = 5$	$M = 10$	$M = 20$	$M = 5$	$M = 10$	$M = 20$
Bias	F.	1,367,167	-2,126	3,386	3,308	3,325	3,231	3,153	3,109
	M.	1,324,310	2,126	-3,386	-3,308	-3,325	-3,231	-3,153	-3,109
RMSE	F.	1,367,167	2,154	6,008	5,888	5,760	5,914	5,637	5,512
	M.	1,324,310	2,154	6,008	5,888	5,760	5,914	5,637	5,512

Note: "F." is "Female" and "M." is "Male".

In Table 4.2, the simulation results can be found that cover the univariate marginal frequencies of the imputed latent variable "Type of family nucleus" in terms of bias and RMSE. Here, the results are very different from the results we found for "Gender", the bias obtained for $Y_{2,1}$ is much higher compared to the bias obtained using MILC under all conditions and the same holds for RMSE. In addition, whether the results for the MILC method depend on the missingness mechanism differ per category. In terms of bias and RMSE, this is the case for the categories "N.A." and "Partners".

Table 4.2
Results in terms of bias and root mean squared error for the four observed categories of the imputed latent variable "Type of family nucleus"

	Type of family nucleus	Frequency	$Y_{2,1}$	MCAR			MAR		
				$M = 5$	$M = 10$	$M = 20$	$M = 5$	$M = 10$	$M = 20$
Bias	Lone parents	97,360	2,670	185	182	176	224	226	220
	N.A.	604,032	8,985	-957	-975	-989	-1,601	-1,612	-1,611
	Partners	1,272,339	-19,686	401	411	427	932	935	932
	Sons/daughters	717,746	8,030	371	381	386	446	451	459
RMSE	Lone parents	97,360	2,672	425	408	395	426	421	414
	N.A.	604,032	8,989	1,337	1,318	1,312	1,837	1,833	1,818
	Partners	1,272,339	19,688	954	914	904	1,256	1,235	1,218
	Sons/daughters	717,746	8,034	630	624	617	715	692	688

Note: "N.A." means "Not applicable". Note that the category "Not stated" is mitigated as it contained zero observations.

In Table 4.3, the simulation results can be found that cover the univariate marginal frequencies of the imputed latent variable "Citizen" in terms of bias and RMSE. Here, the results are comparable to the results we found for "Type of family nucleus", as the bias obtained when only $Y_{3,1}$ is used is again much higher compared to the bias obtained using MILC method and the same holds for RMSE. As was also the case for "Type of family nucleus", whether the results for the MILC method depend on the missingness mechanism differ per category, although this is more the case for the bias here, and not so much in terms of RMSE.

Table 4.3
Results in terms of bias and root mean squared error for the four observed categories of the imputed latent variable "Citizen"

	Citizen	Frequency	$Y_{3,1}$	MCAR			MAR		
				$M = 5$	$M = 10$	$M = 20$	$M = 5$	$M = 10$	$M = 20$
Bias	EU	79,212	51,365	-5	-7	-12	-199	-211	-216
	NL	2,511,214	-116,899	-555	-546	-545	117	124	107
	not EU	89,592	58,085	512	502	507	62	69	89
	Not stated	11,459	7,448	49	51	49	21	18	20
RMSE	EU	79,212	51,365	410	398	388	488	486	475
	NL	2,511,214	116,899	925	894	883	767	756	720
	not EU	89,592	58,086	800	770	767	618	611	590
	Not stated	11,459	7,449	201	197	190	204	205	198

Note: "N.S." means "Not stated". Note that the category "Stateless" is mitigated as it contained zero observations.

Boeschoten et al. (2017) concluded that the quality of the output when MILC is applied related to how well the latent class model is able to make classifications based on the observed data, which is summarized in the entropy R^2. The entropy R^2 values for "Gender", "Type of family nucleus" and "Citizen" are approximately 0.7352, 0.9191, and 0.8571 respectively under MCAR. So this corresponds to the quality of the results for the latent variables in terms of bias and RMSE. An additional explanation for "Gender" is that the two categories are of comparable size and the amount of misclassification in both categories is approximately equal and behaves symmetrical in our simulation study. This causes that the marginal distribution of $Y_{1,1}$ is very similar to the marginal distribution of X_1 and not so much affected by misclassification.

4.1.2 Joint frequencies of imputed variables

In Table 4.4, the simulation results can be found that cover the joint marginal frequencies of the three imputed latent variables in terms of bias and RMSE. Again, it can be seen here that if only $Y_{v,1}$ is used, severe bias is present in all cells of the joint frequency table. The results obtained when the MILC method is applied show much lower amounts of bias and RMSE. Here, the differences between different numbers for M or different missingness mechanism are much smaller compared to the differences between MILC and $Y_{v,1}$. Furthermore, the differences in the amount of bias for particular cells after applying the MILC method seem to be related to imbalances in cell frequencies within particular variables. More specifically, the variable "Citizen" knows substantive differences in cell frequencies and within Table 4.4, it can be seen that particular the category "not EU" is affected in terms of bias by this imbalance.

Table 4.4
Results in terms of bias and root mean squared error for the 32 observed categories of the joint distribution of the three imputed latent variables "Gender", "Type of family nucleus" and "Citizen"

	Gender × Type of family nucleus × Citizen					MCAR			MAR		
	Gender	Family nucleus	Citizen	Frequency	$Y_{v,1}$	$M = 5$	$M = 10$	$M = 20$	$M = 5$	$M = 10$	$M = 20$
Bias	F.	Lone parents	EU	2,091	1,434	8	7	7	1	0	0
	F.	Lone parents	NL	76,131	-6,620	652	650	646	240	241	234
	F.	Lone parents	not EU	3,120	1,513	33	32	32	39	39	38
	F.	Lone parents	N.S.	646	154	-5	-5	-6	-13	-13	-13
	F.	N.A.	EU	12,436	5,971	433	432	432	431	427	427
	F.	N.A.	NL	293,960	-11,998	-595	-618	-623	905	891	880
	F.	N.A.	not EU	9,509	7,317	1,032	1,031	1,032	1,069	1,069	1,071
	F.	N.A.	N.S.	1,221	982	182	182	182	198	197	197
	F.	Partners	EU	20,443	11,185	237	236	235	24	19	21
	F.	Partners	NL	584,547	-34,001	294	262	279	-564	-599	-624
	F.	Partners	not EU	26,877	12,022	404	402	401	254	255	258
	F.	Partners	N.S.	1,292	1,837	-19	-18	-18	-23	-24	-24
	F.	Sons/daughters	EU	4,368	7,541	-778	-779	-780	-851	-853	-854
	F.	Sons/daughters	NL	321,364	-8,738	2,483	2,471	2,479	2,620	2,601	2,588
	F.	Sons/daughters	not EU	7,680	8,303	-764	-768	-766	-876	-874	-869
	F.	Sons/daughters	N.S.	1,482	971	-209	-208	-208	-223	-223	-222
	M.	Lone parents	EU	389	591	-10	-11	-11	9	9	9
	M.	Lone parents	NL	14,536	4,791	-553	-552	-554	-134	-131	-130
	M.	Lone parents	not EU	372	707	35	35	35	53	53	53
	M.	Lone parents	N.S.	75	100	27	27	27	28	29	29
	M.	N.A.	EU	16,308	4,444	-306	-304	-305	-349	-349	-350
	M.	N.A.	NL	253,493	-3,733	-714	-708	-717	-2,730	-2,722	-2,713
	M.	N.A.	not EU	13,636	5,548	-904	-903	-902	-1,023	-1,023	-1,020
	M.	N.A.	N.S.	3,469	455	-85	-86	-87	-102	-103	-104
	M.	Partners	EU	18,444	11,881	793	796	794	905	906	906
	M.	Partners	NL	599,278	-38,164	-3,170	-3,128	-3,127	-1,528	-1,490	-1,474
	M.	Partners	not EU	19,776	13,709	1,794	1,793	1,793	1,785	1,790	1,791
	M.	Partners	N.S.	1,682	1,846	69	69	69	78	78	79
	M.	Sons/daughters	EU	4,733	8,319	-382	-382	-384	-370	-371	-374
	M.	Sons/daughters	NL	367,905	-18,435	1,049	1,076	1,072	1,308	1,333	1,346
	M.	Sons/daughters	not EU	8,622	8,966	-1,118	-1,120	-1,117	-1,240	-1,239	-1,233
	M.	Sons/daughters	N.S.	1,592	1,103	90	90	91	77	77	78

Note: "N.S." means "Not stated" and "N.A." means "Not applicable". Note that the categories "Stateless" for "Citizen" and "Not Stated" for "Type of family nucleus" are mitigated as they contained zero observations.

Table 4.4 (continued)
Results in terms of bias and root mean squared error for the 32 observed categories of the joint distribution of the three imputed latent variables "Gender", "Type of family nucleus" and "Citizen"

	Gender × Type of family nucleus × Citizen					MCAR			MAR		
	Gender	Family nucleus	Citizen	Frequency	$Y_{v,1}$	$M = 5$	$M = 10$	$M = 20$	$M = 5$	$M = 10$	$M = 20$
RMSE	F.	Lone parents	EU	2,091	1,434	45	42	41	45	42	40
	F.	Lone parents	NL	76,131	6,621	742	734	724	418	408	394
	F.	Lone parents	not EU	3,120	1,514	67	64	64	71	68	66
	F.	Lone parents	N.S.	646	155	22	21	20	26	25	24
	F.	N.A.	EU	12,436	5,972	449	446	445	447	442	440
	F.	N.A.	NL	293,960	12,001	1,260	1,245	1,222	1,433	1,374	1,348
	F.	N.A.	not EU	9,509	7,317	1,038	1,037	1,037	1,075	1,075	1,076
	F.	N.A.	N.S.	1,221	983	185	185	185	202	201	201
	F.	Partners	EU	20,443	11,186	291	285	282	173	163	157
	F.	Partners	NL	584,547	34,003	2,332	2,285	2,204	2,364	2,248	2,197
	F.	Partners	not EU	26,877	12,023	456	450	447	330	327	327
	F.	Partners	N.S.	1,292	1,838	46	44	43	48	48	47
	F.	Sons/daughters	EU	4,368	7,541	787	787	787	860	862	863
	F.	Sons/daughters	NL	321,364	8,742	2,820	2,796	2,781	2,959	2,903	2,879
	F.	Sons/daughters	not EU	7,680	8,304	779	782	780	892	889	883
	F.	Sons/daughters	N.S.	1,482	972	216	214	214	230	230	229
	M.	Lone parents	EU	389	592	18	17	17	17	17	16
	M.	Lone parents	NL	14,536	4,792	605	600	600	271	260	257
	M.	Lone parents	not EU	372	707	38	38	37	55	55	55
	M.	Lone parents	N.S.	75	101	27	27	27	29	29	29
	M.	N.A.	EU	16,308	4,445	331	328	327	373	371	370
	M.	N.A.	NL	253,493	3,742	1,390	1,349	1,314	2,959	2,931	2,911
	M.	N.A.	not EU	13,636	5,549	913	912	911	1,033	1,031	1,028
	M.	N.A.	N.S.	3,469	456	107	105	104	121	121	120
	M.	Partners	EU	18,444	11,881	808	810	807	919	919	917
	M.	Partners	NL	599,278	38,165	3,898	3,837	3,794	2,755	2,617	2,568
	M.	Partners	not EU	19,776	13,709	1,804	1,803	1,803	1,797	1,800	1,800
	M.	Partners	N.S.	1,682	1,846	88	87	85	98	95	95
	M.	Sons/daughters	EU	4,733	8,319	403	403	403	401	401	402
	M.	Sons/daughters	NL	367,905	18,437	1,728	1,723	1,687	1,905	1,872	1,854
	M.	Sons/daughters	not EU	8,622	8,967	1,129	1,130	1,127	1,252	1,250	1,244
	M.	Sons/daughters	N.S.	1,592	1,104	109	108	107	103	102	101

Note: "N.S." means "Not stated" and "N.A." means "Not applicable". Note that the categories "Stateless" for "Citizen" and "Not Stated" for "Type of family nucleus" are mitigated as they contained zero observations.

4.1.3 Restricted cells

In Table 4.5, the simulation results can be found for the six cells that are restricted in the marginal cross-table between "Age" and "Type of family nucleus". Under "Frequency", it can be seen that these six cells should all contain zero observations. A combination of these scores is logically impossible. Furthermore, it can be seen that due to misclassification in $Y_{2,1}$, observations containing these combinations of scores are present when $Y_{2,1}$ is used to estimate this cross-table directly. In addition, it can be seen that if the MILC method is applied, such impossible combinations of scores will never be present, regardless of the missingness mechanism or the number of imputations. Furthermore, as the cells in this marginal table contain zero observations, all cells of more detailed tables covering these logically impossible combinations of scores automatically also contain zero observations.

Table 4.5
Results in terms of bias and root mean squared error for the six restricted categories from cross-table between "Type of family nucleus" and the covariate "Age"

	Type of family nucleus		Frequency	$Y_{2,1}$	MCAR			MAR		
					$M = 5$	$M = 10$	$M = 20$	$M = 5$	$M = 10$	$M = 20$
Bias	Lone parents	under 5 years	0	377	0	0	0	0	0	0
	Lone parents	5 to 9 years	0	386	0	0	0	0	0	0
	Lone parents	10 to 14 years	0	376	0	0	0	0	0	0
	Partners	under 5 years	0	4,934	0	0	0	0	0	0
	Partners	5 to 9 years	0	5,041	0	0	0	0	0	0
	Partners	10 to 14 years	0	4,937	0	0	0	0	0	0
RMSE	Lone parents	under 5 years	0	377	0	0	0	0	0	0
	Lone parents	5 to 9 years	0	386	0	0	0	0	0	0
	Lone parents	10 to 14 years	0	377	0	0	0	0	0	0
	Partners	under 5 years	0	4,934	0	0	0	0	0	0
	Partners	5 to 9 years	0	5,041	0	0	0	0	0	0
	Partners	10 to 14 years	0	4,937	0	0	0	0	0	0

4.1.4 The complete population frequency table

Figures 4.1 and 4.2 show results in terms of bias and root mean squared error (RMSE) when the complete census table, so the cross-table between the six variables, is estimated. As these are 42,000 cells in total, it is not feasible to evaluate them individually. Figure 4.1 and Figure 4.2 give an overview of how size of the cell frequency is related to the quality of the results. Here it can be seen that if $Y_{v,1}$ are used, results in terms of bias and RMSE are related directly to cell frequency. More specifically, the relationship between cell frequency and absolute bias is approximately linear where the amount of bias is approximately 10% of the cell frequency.

Figure 4.1 Results in terms of bias when the complete cross-table between the latent variables "Gender", "Type of family nucleus" and "Citizen" and the three covariates "Age", "Marital status" and "Place of birth" is estimated. The X-axis represents cell frequency and the Y-axis represents the bias. Results are shown for $Y_{v,1}$, MILC-MCAR-20 and MILC-MAR-20.

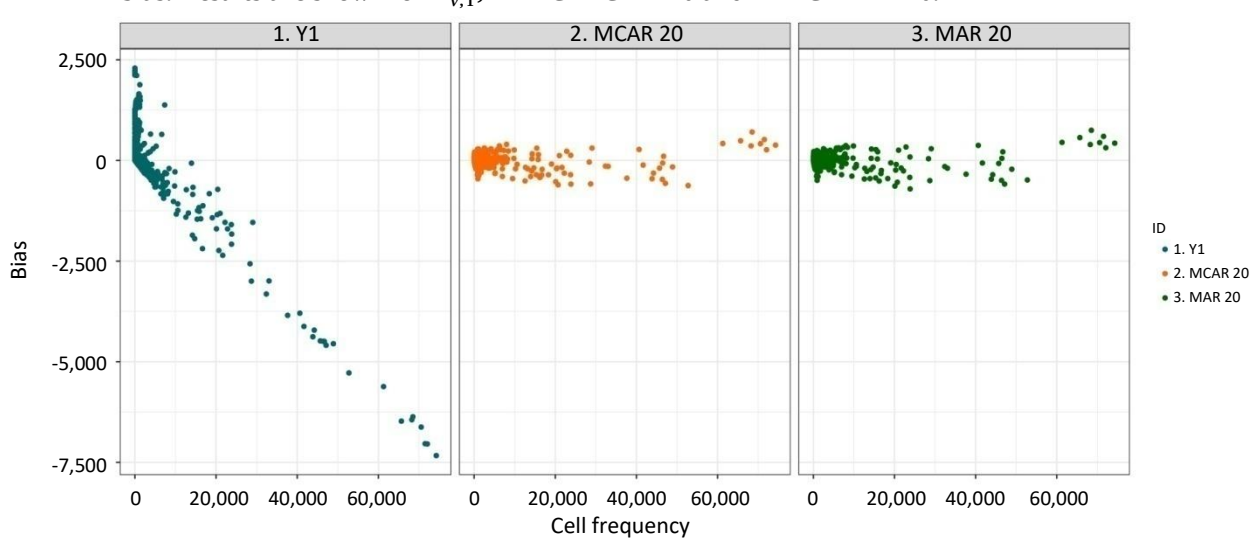
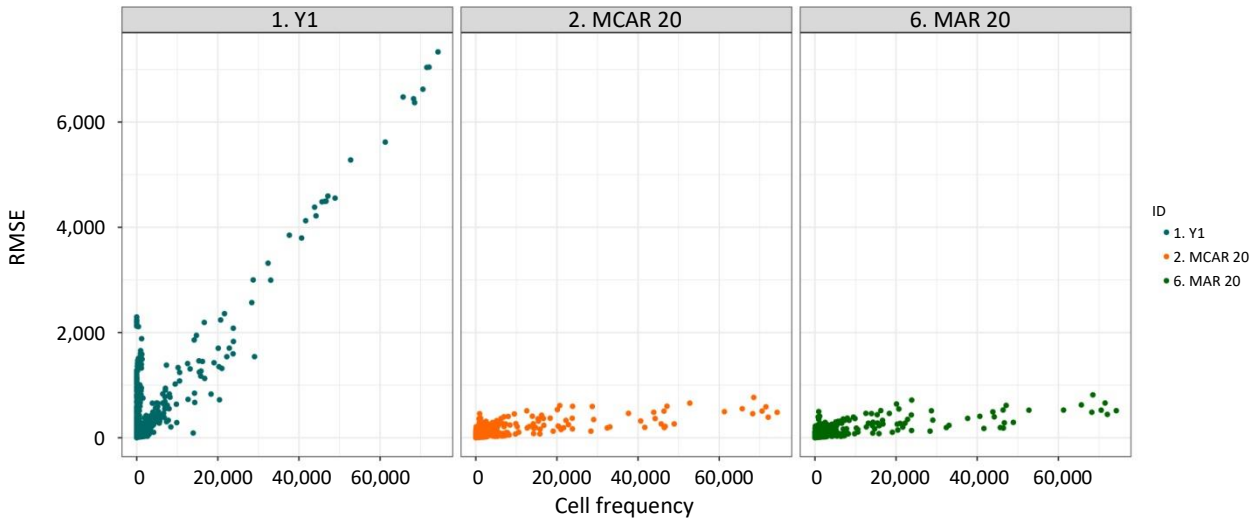


Figure 4.2 Results in terms of root mean squared error (RMSE) when the complete cross-table between the latent variables "Gender", "Type of family nucleus" and "Citizen" and the three covariates "Age", "Marital status" and "Place of birth" is estimated. The X-axis represents cell frequency and the Y-axis represents the RMSE. Results are shown for $Y_{v,1}$, MILC-MCAR-20 and MILC-MAR-20.



4.2 Results in terms of variance

4.2.1 Univariate marginal frequencies of imputed variables

In Table 4.6, the simulation results can be found that cover the univariate marginal frequencies "Gender" in terms of se/sd. As this ratio measures whether the average standard error estimated at each replication in the simulation correctly describes the uncertainty (standard deviation) that is found over the estimates, it should be close to one. In addition, as a completely observed and finite population is assumed, variance is not estimated when $Y_{v,1}$ is used. The results obtained using MILC are generally close to one and comparable to the results in terms of bias as only minor differences can be found between different values for M or between the different missingness mechanisms.

Table 4.6
Results in terms of average standard error of the estimates divided by standard deviation over the estimates (se/sd) for the two categories of the imputed latent variable "Gender"

	Gender	Frequency	$Y_{v,1}$	MCAR			MAR		
				$M = 5$	$M = 10$	$M = 20$	$M = 5$	$M = 10$	$M = 20$
se/sd	F.	1,367,167	-	1.0540	1.0317	1.0363	1.0030	1.0235	1.0237
	M.	1,324,310	-	1.0546	1.0317	1.0363	1.0034	1.0236	1.0236

Note: ("F." is "Female" and "M." is "Male").

In Table 4.7 and 4.8, the simulation results can be found that cover the univariate marginal frequencies for "Type of family nucleus" and "Citizen" respectively in terms of se/sd. The results found here have a very comparable structure compared to the results we found for "Gender".

Table 4.7
Results in terms of average standard error of the estimates divided by standard deviation over the estimates (se/sd) for the four observed categories of the imputed latent variable "Type of family nucleus"

	Type of family nucleus	Frequency	$Y_{v,1}$	MCAR			MAR		
				$M = 5$	$M = 10$	$M = 20$	$M = 5$	$M = 10$	$M = 20$
se/sd	Lone parents	97,360	-	1.0457	1.0510	1.0529	1.0561	1.0337	1.0336
	N.A.	604,032	-	0.9706	0.9874	0.9922	0.9751	0.9829	0.9863
	Partners	1,272,339	-	1.0332	1.0418	1.0456	1.0052	1.0269	1.0298
	Sons/daughters	717,746	-	0.9594	0.9615	0.9606	0.9696	0.9880	0.9938

Note: "N.A." means "Not applicable". Note that the category "Not stated" is mitigated as it contained zero observations.

Table 4.8
Results in terms of average standard error of the estimates divided by standard deviation over the estimates for the four observed categories of the imputed latent variable "Citizen"

	Type of family nucleus	Frequency	$Y_{v,1}$	MCAR			MAR		
				$M = 5$	$M = 10$	$M = 20$	$M = 5$	$M = 10$	$M = 20$
se/sd	Citizen EU	79,212	-	1.0417	1.0172	1.0362	1.0768	1.0539	1.0571
	Citizen NL	2,511,214	-	1.0136	1.0113	1.0235	1.0925	1.0645	1.0927
	Citizen not EU	89,592	-	0.9478	0.9632	0.9709	1.0282	0.9916	1.0125
	Not stated	11,459	-	1.0063	1.0208	1.0238	1.1057	1.0861	1.1143

Note: "N.S." means "Not stated". Note that the category "Stateless" is mitigated as it contained zero observations.

4.2.2 Joint frequencies of imputed variables

In Table 4.9, the simulation results can be found that cover the joint marginal frequencies of the imputed latent variables "Gender", "Type of family nucleus" and "Citizen" in terms of absolute se/sd. The results found for these joint frequencies are very comparable to the results we found for the marginal frequencies. For cells with a relatively low frequency, it can be seen that the ratio is in general larger than one, indicating that the variance estimated for these frequencies (and thereby the differences between the imputations) incorporate more uncertainty than is actually found over different replications. Summarizing, the uncertainty for cells containing low frequencies is overestimated.

Results in terms for variance are not shown for the restricted cells, as a variance term cannot be estimated here.

Table 4.9
Results in terms of average standard error of the estimates divided by standard deviation over the estimates for the 32 observed categories of the joint distribution of the three imputed latent variables "Gender", "Type of family nucleus" and "Citizen"

Gender × Type of family nucleus × Citizen			Frequency	$Y_{v,1}$	MCAR			MAR		
Gender	Family nucleus	Citizen			$M = 5$	$M = 10$	$M = 20$	$M = 5$	$M = 10$	$M = 20$
F.	Lone parents	EU	2,091	-	1.1813	1.2097	1.2032	1.1495	1.1654	1.1997
F.	Lone parents	NL	76,131	-	1.0371	1.0471	1.0504	1.0270	1.0252	1.0349
F.	Lone parents	not EU	3,120	-	1.1659	1.1590	1.1519	1.1607	1.1634	1.1870

Note: "N.S." means "Not stated" and "N.A." means "Not applicable". Note that the categories "Stateless" for "Citizen" and "Not Stated" for "Type of family nucleus" are mitigated as they contained zero observations.

Table 4.9 (continued)
Results in terms of average standard error of the estimates divided by standard deviation over the estimates for the 32 observed categories of the joint distribution of the three imputed latent variables "Gender", "Type of family nucleus" and "Citizen"

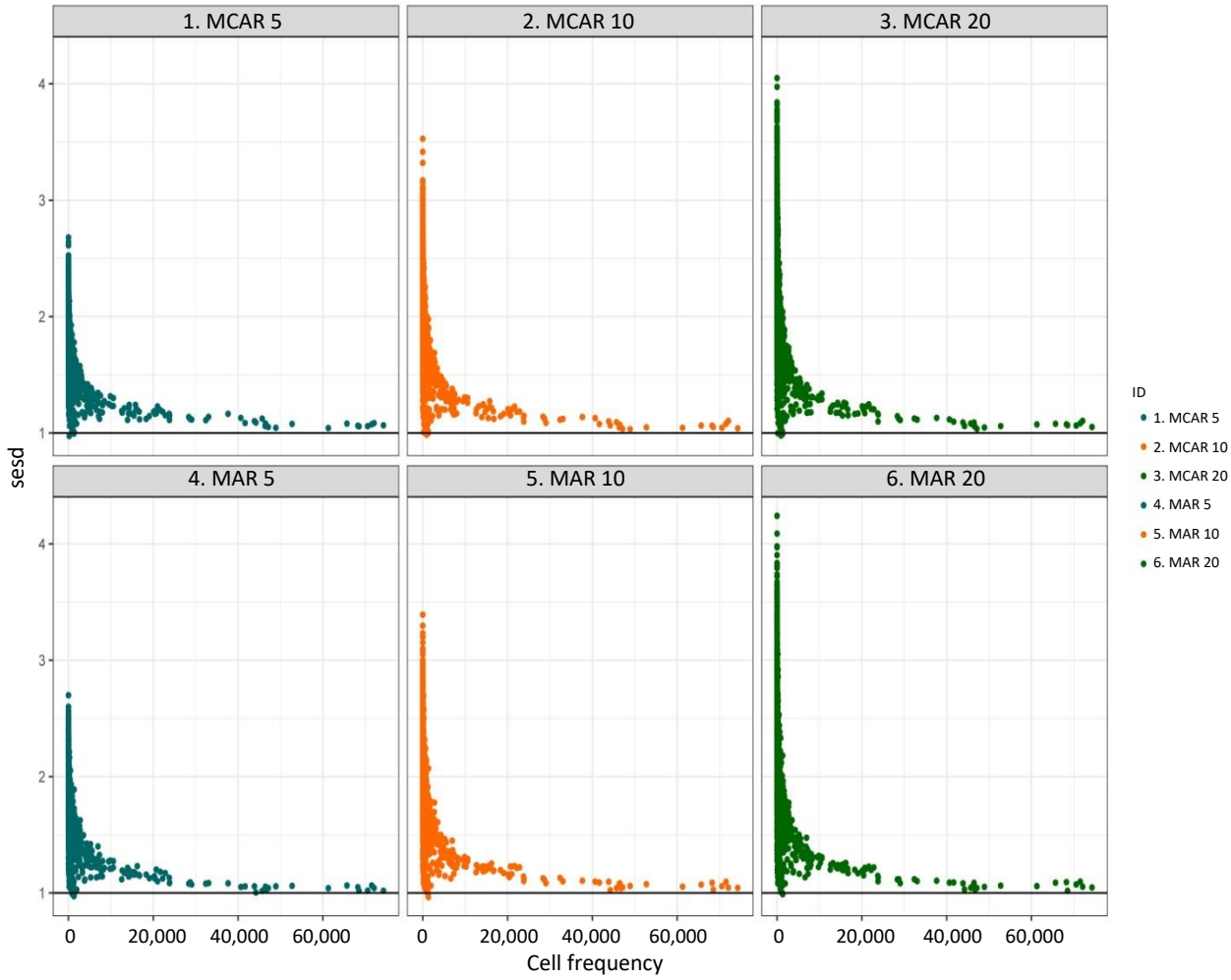
Gender × Type of family nucleus × Citizen					MCAR			MAR		
Gender	**Family nucleus**	**Citizen**	**Frequency**	$Y_{v,1}$	$M = 5$	$M = 10$	$M = 20$	$M = 5$	$M = 10$	$M = 20$
F.	Lone parents	N.S.	646	-	1.0963	1.1004	1.1272	1.1110	1.1000	1.1054
F.	N.A.	EU	12,436	-	1.0850	1.0838	1.1172	1.0888	1.1065	1.1456
F.	N.A.	NL	293,960	-	1.0840	1.0652	1.0575	1.0158	1.0406	1.0461
F.	N.A.	not EU	9,509	-	1.1636	1.1822	1.1892	1.1574	1.1383	1.1562
F.	N.A.	N.S.	1,221	-	1.1789	1.1964	1.2097	1.1959	1.1826	1.2133
F.	Partners	EU	20,443	-	1.0508	1.0537	1.0653	1.0689	1.0684	1.0925
F.	Partners	NL	584,547	-	1.0313	1.0099	1.0189	1.0035	1.0253	1.0197
F.	Partners	not EU	26,877	-	1.0532	1.0766	1.0720	1.0765	1.0725	1.0733
F.	Partners	N.S.	1,292	-	1.1471	1.1566	1.1504	1.2157	1.1855	1.1940
F.	Sons/daughters	EU	4,368	-	1.0135	1.0147	1.0338	1.0430	1.0518	1.0479
F.	Sons/daughters	NL	321,364	-	1.0548	1.0379	1.0527	1.0017	1.0222	1.0221
F.	Sons/daughters	not EU	7,680	-	0.9977	0.9966	0.9909	1.0249	1.0132	1.0416
F.	Sons/daughters	N.S.	1,482	-	1.0344	1.0325	1.0357	1.0836	1.0688	1.0890
M.	Lone parents	EU	389	-	1.3198	1.4136	1.4316	1.2941	1.3575	1.4470
M.	Lone parents	NL	14,536	-	1.0784	1.0762	1.0736	1.0755	1.0690	1.0650
M.	Lone parents	not EU	372	-	1.4159	1.3857	1.4511	1.4814	1.4481	1.4619
M.	Lone parents	N.S.	75	-	1.4330	1.5192	1.5659	1.4598	1.5035	1.5373
M.	N.A.	EU	16,308	-	1.0990	1.0908	1.1165	1.0894	1.1022	1.1366
M.	N.A.	NL	253,493	-	1.0035	1.0100	1.0193	0.9920	1.0175	1.0238
M.	N.A.	not EU	13,636	-	1.1168	1.1100	1.1141	1.0826	1.1054	1.0952
M.	N.A.	N.S.	3,469	-	1.0241	1.0818	1.1052	1.1592	1.1478	1.1780
M.	Partners	EU	18,444	-	1.1618	1.1593	1.1579	1.1473	1.1335	1.1476
M.	Partners	NL	599,278	-	1.0668	1.0444	1.0487	1.0081	1.0329	1.0231
F.	Partners	not EU	19,776	-	1.0932	1.0788	1.0816	1.0674	1.0612	1.0911
F.	Partners	N.S.	1,682	-	1.1068	1.1411	1.1418	1.1335	1.1719	1.1770
F.	Sons/daughters	EU	4,733	-	1.0598	1.0396	1.0548	1.0528	1.0497	1.0414
F.	Sons/daughters	NL	367,905	-	1.0549	1.0347	1.0365	1.0098	1.0298	1.0340
F.	Sons/daughters	not EU	8,622	-	1.0077	1.0093	1.0100	1.0413	1.0449	1.0471
F.	Sons/daughters	N.S.	1,592	-	1.0472	1.0617	1.0699	1.0458	1.0362	1.0627

Note: ("N.S." means "Not stated" and "N.A." means "Not applicable"). Note that the categories "Stateless" for "Citizen" and "Not Stated" for "Type of family nucleus" are mitigated as they contained zero observations.

4.2.3 The complete population frequency table

In Figure 4.3, results can be found in terms of average standard error of the cell frequencies divided by the standard deviation over the frequencies estimated in the 500 replications in the simulation study (se/sd). Here it can be seen that the standard error estimated per cell frequency is especially too large when cell frequencies are close to zero, and become closer to the nominal rate of one as the cell frequencies become larger. Apparently, variability due to missing and conflicting values is overestimated by MILC for cells with a frequency close to zero. In addition, this becomes more apparent when the number of imputations increases and it is not influenced by missingness mechanism.

Figure 4.3 Results in terms of average standard error of the cell frequencies divided by the standard deviation over the frequencies (se/sd) when the complete cross-table between the latent variables "Gender", "Type of family nucleus" and "Citizen" and the three covariates "Age", "Marital status" and "Place of birth" is estimated. The X-axis represents cell frequency and the Y-axis represents the se/sd ratio. Results are shown for MILC-MCAR-5, MILC-MCAR-10, MILC-MCAR-20, MILC-MAR-5, MILC-MAR-10 and MILC-MAR-20.



4.3 Sensitivity to violations of assumptions

The simulation study presented in this paper is aimed at investigating the performance of the MILC method in a situation of misclassification in a finite population setting. When applying the MILC method in practice, a number of assumptions are made and during this simulation study these assumptions were met. To further investigate the sensitivity to violations of these assumptions, additional simulation studies were performed.

An important assumption made when applying the MILC method is that the missingness mechanism is either MCAR or MAR. Therefore, a first sensitivity analysis involves a Missing Not At Random (MNAR) mechanism. More specifically, we generated this mechanism in such a way that the probability of being

missing in the survey indicator for "Type of family nucleus" depends on the latent variable "type of family nucleus" and is smallest for the first category and largest for the last category. In Table 4.10, it can be seen that the bias and RMSE increase when the mechanism is MNAR compared to MAR, while the se/sd is not affected. More specifically, it can be seen that the extent of the bias relates to how much the respective class is affected by the mechanism.

A second assumption states that the measurement error present in the indicators is random. To investigate sensitivity to the violation of this assumption, we generated a selective measurement error mechanism where the probability of measurement error in the register indicator for the variable "type of family nucleus" differs per category. Here, again the first category is least affected and the last category most. In Table 4.10 it can be seen that the effect of this selective mechanism are limited. The bias increases in a similar way as the percentage of measurement error in the respective category increases, but these are still relatively low amounts of bias. The se/sd is not affected by the mechanism.

Table 4.10
Results in terms of bias, root mean squared error and se/sd for the four observed categories of the imputed latent variable "Type of family nucleus"

	Type of family nucleus	Frequency	$Y_{2,1}$	MAR	MNAR	Selective	ME covar
Bias	Lone parents	97,360	2,670	224	6,256	105	1,172,993
	N.A.	604,032	8,985	-1,601	27,002	-1,824	534
	Partners	1,272,339	-19,686	932	-11,341	1,116	-1,174,697
	Sons/daughters	717,746	8,030	446	-21,917	603	1,170
RMSE	Lone parents	97,360	2,672	426	6,268	332	1,172,994
	N.A.	604,032	8,989	1,837	27,017	2,060	1,094
	Partners	1,272,339	19,688	1,256	11,377	1,466	1,174,697
	Sons/daughters	717,746	8,034	715	21,924	819	1,291
se/sd	Lone parents	97,360	-	1.0561	1.01936	1.0634	1.0518
	N.A.	604,032	-	0.9751	1.02491	0.9722	1.0471
	Partners	1,272,339	-	1.0052	0.97456	0.9291	0.9649
	Sons/daughters	717,746	-	0.9696	1.02547	1.0962	1.0181

Note: "N.A." means "Not applicable" under different violations of assumptions. Note that the category "Not stated" is mitigated as it contained zero observations.

A third assumption is that covariates do not contain measurement error. This assumption is the most remarkable, as it is typically often not the case that a coviarate does not contain measurement error. It is more likely that these variables will be treated as such because no additional information about their measurement error is known. If information was known, for example because additional survey information was present, it would have been incorporated by means of a latent variable. As in practice however there is always a probability that for some variables such information is not known, we investigate the sensitivity of the method to violation of this assumption. More specifically, we generated 5% misclassification in the covariate "marital status", which has a relatively strong association with the latent variable "type of family nucleus". Indeed, the bias in some categories is highly affected by this misclassification.

5. Discussion

In this paper, the performance of the MILC method was investigated in a situation where misclassification was induced in a finite population setting. Here, an existing population census table was used as a starting point, and for three categorical variables present in this census table, two indicator variables were generated with 5% misclassification each, where one indicator also contains approximately 90% missing values. As a finite population was assumed, the estimated variance only contained a between variance component reflecting the differences between the imputations and thereby the uncertainty caused by the misclassification and missing values in the indicator variables.

The simulation results show that the method, regardless of the number of imputations, produces results with a low bias for marginal frequency distributions, cross-tables between imputed latent variables and covariates and even for the complete six-way cross-table. Striking is the amount of bias that is induced when the indicator observed via the register is used to calculate the cross-tables evaluated in comparison to when MILC is used. It is also shown that if these indicators are used, it is likely that impossible combinations of scores are produced as well, something that can be easily circumvented by specifying edit restrictions in the LC model. This simulation study once again shows that misclassification, even if it is non-systematic, can seriously bias results. In terms of variance, it was seen that if the MILC method is applied, variance estimates are appropriate in general. However, if cell frequencies are relatively small, the variance is overestimated. This problem is more severe if the complete frequency table is evaluated, because this large table contains many cells with low frequencies.

The current set-up of this simulation study knows two major limitations. The first is caused by the large amount of cells in the cross-table. Because of this, a latent class model containing only main effects was used. It was not feasible to use a saturated model as the number of parameters would be very large, and it would be likely that not every parameter is estimable in every bootstrap sample. This would limit the use of starting values, thereby increasing the computation time for the simulation study to an unfeasible amount.

A second limitation is that in our simulation set-up we only considered relatively simple sampling designs for the survey data: simple random sampling (MCAR conditions) and, essentially, stratified simple random sampling (MAR conditions). A future study could examine to what extent the MILC method can also correct for misclassification error with appropriate variance estimates when survey data are obtained by a complex sampling design that involves, for instance, cluster sampling, multistage sampling or sampling with unequal probabilities proportional to size. In the context of missing data it has been found that, although a generally accepted theory is still lacking, in practice multiple imputation often works reasonably well for complex samples, provided that design variables and/or survey weights are included in the imputation model; see, e.g., Rässler (2004, page 14) and the references listed there. It would be interesting to investigate whether this result also applies to multiple imputation in the context of correcting for measurement errors. As an alternative, Zhou, Elliott and Raghunathan (2016) proposed a Bayesian approach to incorporate survey design features into a multiple-imputation analysis.

The starting point of this simulation study was an existing population census table. A nice property here was that we could approach this as a finite and known population. Therefore, we did not have to include (within) sampling variance in our estimate of the total variance. It was insightful to evaluate cell frequencies of both univariate and multivariate cross-tables as results generally appeared to be related to cell-frequency.

References

Bakker, B. (2010). *Micro-integration, state of the art*. Paper presented at the joint UNECE-Eurostat expert group meeting on registered based censuses in The Hague, May 11, 2010. Retrieved from https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.41/2010/wp.10.e.pdf (date visited 2017.04.24).

Bakker, B. (2011). Micro-integration. statistical methods 201108. *Statistics Netherlands*.

Bakker, B., Van Rooijen, J. and Van Toor, L. (2014). The system of social statistical datasets of statistics netherlands: An integral approach to the production of register-based social statistics. *Statistical Journal of the IAOS*, 30(4), 411-424.

Bikker, R., Daalmans, J. and Mushkudiani, N. (2013). Benchmarking large accounting frameworks: A generalized multivariate model. *Economic Systems Research*, 25(4), 390-408.

Boeschoten, L., de Waal, T., and Vermunt, J.K. (2019). Estimating the number of serious road injuries per vehicle type in the netherlands by using multiple imputation of latent classes. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182(4), 1463-1486. Retrieved from https://doi.org/10.1111/rssa.12471.

Boeschoten, L., Filipponi, D. and Varriale, R. (2021). Combining multiple imputation and hidden markov modeling to obtain consistent estimates of employment status. *Journal of Survey Statistics and Methodology*, 9(3), 549-573. Retrieved from https://doi.org/10.1093/jssam/smz052.

Boeschoten, L., Oberski, D. and de Waal, T. (2017). Estimating classification errors under edit restrictions in composite survey-register data using Multiple Imputation Latent Class modelling (MILC). *Journal of Official Statistics*, 33(4), 921-962. Retrieved from https://doi.org/10.1515/jos-2017-0044.

Census Hub (2017, July). *European Statistical System*. Online, July 2017, (last visited 11/07/2017).

Daalmans, J. (2018). Divide-and-Conquer solutions for estimating large consistent table sets. *Statistical Journal of the IAOS*, 34(2), 223-233.

Daalmans, J. (2019). Pushing the Boundaries for Automated Data Reconciliation in Official Statistics. Tilburg University.

de Waal, T., Pannekoek, J. and Scholtus, S. (2011). *Handbook of Statistical Data Editing and Imputation*, New York: John Wiley & Sons, Inc., 563. (ISBN 0470904836, 9780470904831).

de Waal, T., van Delden, A. and Scholtus, S. (2020). Multi-source statistics: Basic situations and methods. *International Statistical Review*, 88(1), 203-228.

Di Fonzo, T., and Martini, M. (2003). Benchmarking systems of seasonally adjusted time series according.

European Commission (2008). Regulation (ec) no 763/2008 of the european parliament and of the council of 9 july 2008 on population and housing censuses. *Official Journal of the European Union*, (L218), 14-20.

European Commission (2009). Commission regulation (ec) no 1201/2009 of 30 november 2009 implementing regulation (ec) no 763/2008 of the european parliament and of the council on population and housing censuses as regards the technical specifications of the topics and of their breakdowns. *Official Journal of the European Union*, (L329), 29-68.

European Commission (2010). Commission regulation (eu) no 1151/2010 of 8 december 2010 implementing regulation (ec) no 763/2008 of the european parliament and of the council on population and housing censuses, as regards the modalities and structure of the quality reports and the technical format for data transmission. *Official Journal of the European Union*, (L324), 1-12.

Geerdinck, M., Goedhuys-van der Linden, M., Hoogbruin, E., De Rijk, A., Sluiter, N. and Verkleij, C. (2014). *Monitor Kwaliteit Stelsel Van Basisregistraties: Nulmeting Van de Kwaliteit Van Basisregistraties in Samenhang, 2014* (13114th Ed.). Henri Faasdreef 312, 2492 JP Den Haag, Centraal Bureau voor de Statistiek. Retrieved from https://www.cbs.nl/-/media/pdf/2016/50/monitor-kwaliteit-stelsel-van-basisregistraties.pdf (date visited 2017.04.25).

Magnus, J.R., van Tongeren, J.W. and de Vos, A.F. (2000). National accounts estimation using indicator ratios. *Review of Income and Wealth*, 46(3), 329-350.

Mashreghi, Z., Haziza, D. and Léger, C. (2016). A survey of bootstrap methods in finite population sampling. *Statistics Surveys*, 10, 1-52.

Pankowska, P., Pavlopoulos, D., Bakker, B. and Oberski, D.L. (2020). Reconciliation of inconsistent data sources using hidden markov models. *Statistical Journal of the IAOS*, 36(4), 1261-1279.

Rässler, S. (2004). The impact of multiple imputation for DACSEIS. (DACSEIS Research Paper Series No. 5).

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63(3), 581-592.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc., 81. Retrieved from dx.doi.org//10.1002/9780470316696 (ISBN 9780471087052) doi: 10.1002/9780470316696.

Särndal, C.-E., Swensson, B. and Wretman, J. (2003). *Model Assisted Survey Sampling*. Springer Science & Business Media.

Schulte Nordholt, E., Van Zeijl, J. and Hoeksma, L. (2014). Dutch census 2011, analysis and methodology. *Statistics Netherlands*. Retrieved from https://www.cbs.nl/-/media/imported/documents/2014/44/2014-b57-pub.pdf (date visited 2017.04.25).

Sefton, J., and Weale, M. (1995). Reconciliation of National Income and Expenditure: Balanced Estimates of National Income for the United Kingdom, 1920-1990. Cambridge University Press, 7.

Stone, R., Champernowne, D.G. and Meade, J.E. (1942). The precision of national income estimates. *The Review of Economic Studies*, 9(2), 111-125.

The Economic and Social Council (2005). Ecosoc resolution 2005/13. 2010 World Population and Housing Census Programme. doi: http://www.un.org/en/ecosoc/docs/2005/resolution%202005-13.pdf.

van Rooijen, J., Bloemendal, C. and Krol, N. (2016). The added value of micro-integration: Data on laid-off employees. *Statistical Journal of the IAOS*, 32(4), 685-692.

Vermunt, J.K., and Magidson, J. (2013a). Latent GOLD 5.0 Upgrade Manual [Computer software manual]. Belmont, MA, Retrieved from https://www.statisticalinnovations.com/wp-content/uploads/LG5manual.pdf (date visited 2017.04.25).

Vermunt, J.K., and Magidson, J. (2013b). Technical guide for *Latent GOLD 5.0: Basic, Advanced, and Syntax*. Statistical Innovations Inc., Belmont, MA. Retrieved from https://www.statisticalinnovations.com/wp-content/uploads/LGtecnical.pdf (date visited 2017.04.25).

Vink, G., and van Buuren, S. (2014). Pooling multiple imputations when the sample happens to be the population. *arXiv* preprint arXiv:1409.8542, Retrieved from https://arxiv.org/abs/1409.8542.

Zhou, H., Elliott, M.R. and Raghunathan, T.E. (2016). A two-step semiparametric method to accommodate sampling weights in multiple imputation. *Biometrics*, 72, 242-252. doi: 10.1111/biom.12413.

Survey Methodology, June 2022
Vol. 48, No. 1, pp. 145-176
Statistics Canada, Catalogue No. 12-001-X

Bayesian inference for multinomial data from small areas incorporating uncertainty about order restriction

Xinyu Chen and Balgobin Nandram[1]

Abstract

When the sample size of an area is small, borrowing information from neighbors is a small area estimation technique to provide more reliable estimates. One of the famous models in small area estimation is a multinomial-Dirichlet hierarchical model for multinomial counts. Due to natural characteristics of the data, making unimodal order restriction assumption to parameter spaces is relevant. In our application, body mass index is more likely at an overweight level, which means the unimodal order restriction may be reasonable. The same unimodal order restriction for all areas may be too strong to be true for some cases. To increase flexibility, we add uncertainty to the unimodal order restriction. Each area will have similar unimodal patterns, but not the same. Since the order restriction with uncertainty increases the inference difficulty, we make comparison with the posterior summaries and approximated log-pseudo marginal likelihood.

Key Words: Bayesian computation; Contingency tables; Log-pseudo marginal likelihood; Monte Carlo method; Small areas; Unimodal order restrictions.

1. Introduction

The term "small area" generally refers to a small geographical area such as a county. It can be described as the sub-population of interest in a large sample survey. Sample survey data certainly can be used to derive reliable estimators of totals and means for large areas or domains. However, using the same survey, sample data for small areas are typically small and likely to yield unacceptably large standard errors (Ghosh and Rao, 1994). Considering the cost and feasibility of conducting new sample survey for small areas, there is a growing demand for reliable small area statistics using the current large sample survey. Pooling information from related areas to find more accurate estimates is key in small area estimation (Rao and Molina, 2015).

With the pooling information feature, Bayesian hierarchical models for small area estimation have lots of potential in small area estimation. It automatically incorporates all sources of uncertainty associated with an inference problem; see, for example, Nandram, Erciulescu and Cruze (2019), Trevisani and Torelli (2007), and You and Rao (2002). In the small area context, multinomial Dirichlet models as one of Bayesian hierarchical models have been widely used for modeling categorical data. Maples (2019) propose a pair of Dirichlet-Multinomial small area models to jointly estimate relevant school-aged child population and poverty. Wang, Berg, Zhu, Sun and Demuth (2018) develop a spatial hierarchical model based on the generalized Dirichlet distribution to construct small area estimators of compositional proportions in several mutually exclusive and exhaustive landcover categories. We focus on extensions of the multinomial Dirichlet model. Recently, there are extensive researches considering constrained inference for small area estimation, for example, Wu, Meyer and Opsomer (2016) and Heck and Davis-

1. Xinyu Chen, Worcester Polytechnic Institute, 100 Institute Rd, Worcester, MA 01609. E-mail: xchen7@wpi.edu; Balgobin Nandram, Worcester Polytechnic Institute, 100 Institute Rd, Worcester, MA 01609.

Stober (2019). Nandram (1997) provided a clear discussion about a hierarchical Bayesian approach for taste-testing experiment and appropriate methods for the model. To select the best population, he studied three criteria based on the distribution of random variables representing values on a hedonic scale using the simple tree order. Nandram (1998) pooled data from several multinomial populations using a three-stage multinomial Dirichlet model.

In many statistical problems, it is necessary to take into account the order restrictions of the unknown parameters of interest. Based on the characteristic of data, incorporating order restrictions on cell probabilities of count data can improve the accuracy of estimation. Our major task is to assume the same unimodal order restrictions across areas in the multinomial Dirichlet model. A lot of discussion have been done about the multinomial Dirichlet model with order restrictions.

Sedransk, Monahan and Chiu (1985) described a Bayesian method for estimation of finite population parameters in general population surveys. They added order restrictions to the model to capture the unimodal smoothness relationships among cell probabilities $(p_1, \ldots, p_J)$, such as

$$p_1 \le \ldots \le p_k \ge p_{k+1} \ge \ldots \ge p_J.$$

But their model cannot pooling information among areas and is not intended for small area estimation.

Gelfand, Smith and Lee (1992) provided very-detailed Gibbs sampler structures for Bayesian analysis of constrained parameters. They suggested that a Dirichlet prior should be used for ordered multinomial parameters, such as $p_1 \le p_2 \le \ldots \le p_k \ge \ldots \ge p_J$. They noted that the Gibbs sampler cannot be employed directly when k is unknown and prior $\Pr(k = j) = w_j,\ j = 1, \ldots, K.$ But the marginal posterior for k can be calculated directly, taking the from

$$\Pr(k = j \mid Y) = \frac{C(\beta_1, \ldots, \beta_J, j) w_j \big/ C(\beta_1 + Y_1, \ldots, \beta_J + Y_J, j)}{\sum_{j=1}^{J} C(\beta_1, \ldots, \beta_J, j) w_j \big/ C(\beta_1 + Y_1, \ldots, \beta_J + Y_J, j)},$$

where $C(\ldots)$ are normalization constant of the Dirichlet distribution with order restrictions. They showed Bayesian inference on order parameters can have higher precision. However, their Dirichlet multinomial model with the ordered parameters does not consider stratification and cannot borrow information among areas either.

Nandram and Sedransk (1995) showed the precision of inference about π_{ij}, the proportion of firms in stratum i belonging to SR class j, can be dramatically increased by using Dirichlet multinomial model with appropriate order restrictions on π_{ij}, within stratum i, $R^{(\ell_s)} = \{\pi_{ij} : \pi_{i1} \le \ldots \le \pi_{i\ell_s} \ge \ldots \ge \pi_{iJ}\}$. Their order restriction is complicated due to the stratification. They also consider the case where there is uncertainty about the vector of modal positions L, which can take g possible values, $\ell_1, \ell_2, \ldots, \ell_g, g \le J$. The position probabilities are given below,

$$\Pr(L = \ell_s) = w_s,\ s = 1, 2, \ldots, g, \text{ where } w_s \text{ are specified and } \sum_{s=1}^{g} w_s = 1.$$

They directly applied Monte Carlo integration to estimate the posterior $w_s = \Pr\left(L = \ell_s \mid \mathbf{n}\right)$. Adopting a Bayesian view, they showed that the posterior variances can be dramatically reduced by including order restrictions among π_{ij}, both within and between the strata. However, their model cannot borrow information among strata and their order restriction assumption is totally different from ours.

Nandram, Sedransk and Smith (1997) improved estimation of the age composition of the population of Atlantic cod with the help of order-restricted Bayesian estimation. Their work was inspired by Sedransk, Monahan and Chiu (1985) and Gelfand, Smith and Lee (1992). Let π_{ij} denote as cell probabilities that a fish belong to a length stratum i and an age class j. To simplify the analysis, the likelihood of $\boldsymbol{\pi}$ is

$$\ell\left(\boldsymbol{\pi} \mid \mathbf{n}\right) \propto \prod_{i=1}^{I} \prod_{j=1}^{J} \pi_{ij}^{n_{ij}}.$$

They took independent Dirichlet distributions as prior; that is

$$f\left(\boldsymbol{\pi} \mid \boldsymbol{\alpha}\right) \propto \prod_{i=1}^{I} \prod_{j=1}^{J} \pi_{ij}^{\alpha_{ij}-1},$$

where $\alpha_{ij} > 0$ is a fixed quantity, within stratum i, $\pi_{i1} \le \ldots \le \pi_{ik_i} \ge \ldots \ge \pi_{iu_i}$ for some $k_i \in Z_i$.

In their Atlantic cod study, let $i = 1$ correspond to the stratum with the shortest fish and $j = 1$ correspond to the youngest fish. It is expected that as i increases, the relative values of the $\left\{\pi_{ij} : j \in Z_i\right\}$ will change. The order restrictions are not just within strata, but also among strata, such as

$$\pi_{i1} \le \ldots \le \pi_{it} \ge \ldots \ge \pi_{iK},$$

$$\pi_{j1} \le \ldots \le \pi_{jt^*} \ge \ldots \ge \pi_{jK} \text{ where } i < j \text{ and } t < t^*.$$

They presented uncertainty about both the locations of the modes and the unimodality itself is included as part of the probabilistic specification, as an extension of their work. They considered the case where there is uncertainty about the vector of modal position L,

$$\Pr\left(L = \ell_s\right) = w_s,\ s = 1, 2, \ldots, g.$$

They showed the joint posterior distribution of π and L is

$$f\left(\pi, L = \ell_s \mid n\right) = \frac{w_s C_{\ell_s}\left(\alpha\right) \prod_{i=1}^{I} g_{n_i}\left(\pi_i\right)}{\sum_{s'=1}^{g} w_s' C_{\ell_s'}\left(\alpha\right) \Big/ C_{\ell_s'}\left(\mathbf{n}\right)}.$$

Their order restriction assumption is not the same across strata, which makes their model is different from ours.

Chen and Nandram (2019), which appeared in the Proceedings of the American Statistical Association, proposed a multinomial Dirichlet model with order restrictions. They considered similar unimodal

structure within each area. They showed how to use the Gibbs sampler to sample the posterior distribution. A huge improvement for estimating the cell probabilities has been shown in their model application. Chen and Nandram (2021) have an overview for this type of order-restricted problem for small area estimation. Their overview cover model selection, sampling from posterior distribution, model diagnostics.

We notice the same unimodal order restrictions may not hold for some cases. Incorporating uncertainty about the order restrictions may solve the issue, see Nandram, Sedransk, and Smith (1997). In our work, to increase the model flexibility, we add uncertainty to the unimodal order restriction. Areas have similar unimodal order restrictions on parameters of interest, but not the same modal position. Our order restrictions occur within areas, not across them and the restriction may not be similar across area. They create a difficult problem that will be discussed in the paper.

The article is organized as follows. In Section 2, we present a brief review of the multinomial Dirichlet model and the multinomial Dirichlet model with order restrictions. In Section 3, we incorporate uncertainty about order restriction into the model. We present the estimation method and show how to use the conditional predictive ordinate as Bayesian diagnostics. In Section 4, for illustrative purpose, we show how to analyze the body mass index (BMI) data using the model incorporating uncertainty about order. We demonstrate how much improvement there is under the order restrictions. In Section 5, we also demonstrate that incorporating uncertainty about order restrictions to the model can improve the robustness of the model. Section 6 has a summary and the future work.

2. Hierarchical multinomial Dirichlet

In this section, we present a brief review of Multinomial-Dirichlet model and its extensions with the order restriction. To study the association between bone mineral density and body mass index (BMI) from several U.S. counties, Nandram, Kim and Zhou (2019) provided a clear discussion of the general hierarchical multinomial Dirichlet model and their methodology for small area estimation. Let n_{ij} be the cell counts, which are numbers in each category j for each area i, θ_{ij} be the corresponding cell probabilities, $i = 1, 2, \ldots, I$, $j = 1, 2, \ldots, K$, and the total number for each area i is $n_{i.} = \sum_{j=1}^{K} n_{ij}$. The general hierarchical multinomial Dirichlet model is

$$\mathbf{n}_i \mid \boldsymbol{\theta}_i \overset{\text{ind}}{\sim} \text{Multinomial}\left(\mathbf{n}_{i.}, \boldsymbol{\theta}_i\right), \mathbf{n}_i = \left(n_{i1}, \ldots, n_{iK}\right),$$

$$\boldsymbol{\theta}_i \mid \boldsymbol{\mu}, \tau \overset{\text{ind}}{\sim} \text{Dirichlet}\left(\boldsymbol{\mu}\tau\right), \boldsymbol{\theta}_i = \left(\theta_{i1}, \ldots, \theta_{ik}\right),$$

$$\pi\left(\boldsymbol{\mu}, \tau\right) = \frac{\left(K-1\right)!}{\left(1+\tau\right)^2},$$

where hyper-parameters $\boldsymbol{\mu} = \left(\mu_1, \ldots, \mu_K\right)$, $\mu_j > 0$, $\sum_{j=1}^{K} \mu_j = 1, \tau > 0.$

They suggest the non-information prior which will be easy to reparameterize. Without any prior information, they take $\boldsymbol{\mu}$ and τ to be independent, $E(\theta_{ij}) = \mu_j$, $\sum_{j=1}^{K} \mu_j = 1$. As an interpretation of hyper-parameters, $\boldsymbol{\mu}$ are related to cell means and τ is related to a prior sample size. This model features stratification and hyper-parameters to pool information from different strata together.

This hierarchical multinomial Dirichlet model is a convenient starting point for small area estimation. For convenience, we denote it as M_1 model for the future discussion.

2.1 Hierarchical multinomial Dirichlet model with order restrictions

Chen and Nandram (2019) incorporate the order restriction into the Bayesian hierarchical multinomial Dirichlet model. Letting n_{ij} be the cell counts, θ_{ij} the corresponding cell probabilities, $i = 1, 2, \ldots, I$, $j = 1, 2, \ldots, K$, $\mathbf{n}_{i.} = \sum_{j=1}^{K} n_{ij}$ and we believe the mode of $\boldsymbol{\theta}_i$'s is $\theta_{im}, 1 \le m \le K$.

Specifically, they assume

$$\mathbf{n}_i \mid \boldsymbol{\theta}_i \overset{\text{ind}}{\sim} \text{Multinomial}(\mathbf{n}_{i.}, \boldsymbol{\theta}_i), \quad \boldsymbol{\theta}_i \in C, \quad i = 1, \ldots, I,$$

where $C = \{\boldsymbol{\theta}_i : \theta_{i1} \le \ldots \le \theta_{im} \ge \ldots \ge \theta_{iK}, i = 1, \ldots, I\}$, and assume the modal position m in C is known.

At the second stage they assume

$$\boldsymbol{\theta}_i \mid \boldsymbol{\mu}, \tau \overset{\text{ind}}{\sim} \text{Dirichlet}(\boldsymbol{\mu}\tau), \; i = 1, \ldots, I,$$

$$\pi(\boldsymbol{\mu}, \tau) = \frac{K(m-1)!(K-m)!}{(1+\tau)^2}, \quad \mu_j > 0, \quad \sum_{j=1}^{K} \mu_j = 1, \quad \boldsymbol{\mu} \in C_{\boldsymbol{\mu}}.$$

Since $E(\theta_{ij}) = \mu_j$, $\boldsymbol{\mu}$ should have the same order restriction as $\boldsymbol{\theta}_i$, which is $\boldsymbol{\mu} \in C_{\boldsymbol{\mu}}$,

$$C_{\boldsymbol{\mu}} = \{\boldsymbol{\mu} : \mu_1 \le \ldots \le \mu_m \ge \ldots \ge \mu_K\},$$

and we assume the modal position m in C_μ is known.

A posteriori $\boldsymbol{\theta}_i \mid \boldsymbol{\mu}, \tau, \mathbf{n}_i \overset{\text{ind}}{\sim} \text{Dirichlet}(\mathbf{n}_i + \boldsymbol{\mu}\tau)$, $\boldsymbol{\theta}_i \in C_i, i = 1, \ldots, I$, where

$$f_{\boldsymbol{\theta}_i \mid \boldsymbol{\mu}, \tau, \mathbf{n}} = \frac{\dfrac{\Gamma\left[\sum_{j=1}^{K}(n_{ij} + \mu_j\tau)\right]}{\prod_{j=1}^{K}\Gamma(n_{ij} + \mu_j\tau)} \prod_{j=1}^{K} \theta_{ij}^{n_{ij} + \mu_j\tau - 1}}{C(\mathbf{n}_i + \boldsymbol{\mu}\tau)},$$

where

$$C(\mathbf{n}_i + \boldsymbol{\mu}\tau) = \int_{\boldsymbol{\theta}_i \in C} \frac{\Gamma\left[\sum_{j=1}^{K}(n_{ij} + \mu_j\tau)\right]}{\prod_{j=1}^{K}\Gamma(n_{ij} + \mu_j\tau)} \prod_{j=1}^{K} \theta_{ij}^{n_{ij} + \mu_j\tau - 1} d\boldsymbol{\theta}_i.$$

In our BMI data application, there are five categories of BMI. We only interest in the normal and overweight BMI level. We use model M_2 represent the model with order restrictions and its modal

position is at the second, which is normal weight. Model M_3 represents the model with order restrictions and its modal position is at the third, which is overweight weight. M_2 and M_3 are the same hierarchical multinomial Dirichlet model, but with different order restrictions.

The joint posterior density of M_2 or M_3 is

$$\pi(\boldsymbol{\theta}, \boldsymbol{\mu}, \tau \mid \mathbf{n}) \propto \prod_{i=1}^{I} \left\{ \prod_{j=1}^{K} \theta_{ij}^{n_{ij}} \frac{1}{D(\boldsymbol{\mu}\tau) C(\boldsymbol{\mu}\tau)} \prod_{j=1}^{K} \theta_{ij}^{\mu_j \tau - 1} \right\} \frac{K(m-1)!(K-m)!}{(1+\tau)^2},$$

where

$$D(\boldsymbol{\mu}\tau) = \frac{\prod_{j=1}^{K} \Gamma(\mu_j \tau)}{\Gamma\left(\sum_{j=1}^{K} \mu_j \tau\right)}$$

is the normalization constant of Dirichlet distribution,

$$C(\boldsymbol{\mu}\tau) = \int_{\boldsymbol{\theta}_i \in C} \frac{\Gamma\left(\sum_{j=1}^{K} \mu_j \tau\right)}{\prod_{j=1}^{K} \Gamma(\mu_j \tau)} \prod_{j=1}^{K} \theta_{ij}^{\mu_j \tau - 1} d\boldsymbol{\theta}_i,$$

is the normalization constant of the truncated Dirichlet distribution, $\boldsymbol{\theta} \in C, \boldsymbol{\mu} \in C_\mu$.

Nandram (1998) showed how to generate samples from model M_1. In fact, using the griddy Gibbs sampler, it can be done easier than the method in Nandram (1998). Chen and Nandram (2019) present sampling methods for $\boldsymbol{\mu}$ and $\boldsymbol{\theta}$ with order restrictions from the joint posterior distribution of model M_2 and M_3, as in Appendix A.1 and Appendix A.2.

Gelfand, Dey and Chang (1992) used predictive distributions to address the issues of model adequacy and model selection. They proposed the conditional predictive ordinate for the model determination. The conditional predictive ordinate (CPO) is based on leave-one-out cross validation. CPO estimates the probability of observing n_i in the future if after having already observed $n_{(i)}$. The sum of the log CPO's is an estimator for the log marginal likelihood. The "best" model amongst competing models have the largest LPML.

Chen and Nandram (2021) presented a method to compute the conditional predictive ordinate (CPO) and LPML as a Bayesian model selection criteria. In Appendix A.3, we have improved estimation to integrate out the order-restricted θ, and the estimated CPO of M_2 and M_3 are

$$\widehat{\mathrm{CPO}}_{i\,(M_2 \text{ or } M_3)} = \left[\frac{1}{M} \sum_{h=1}^{M} \frac{\prod_{j=1}^{K} n_{ij}!}{n_{i.}!} \left(\frac{1}{M'} \sum_{h'=1}^{M'} \prod_{j=1}^{K} \theta_{ij}^{(h')^{-n_{ij}}} \right) \right]^{-1},$$

where $\boldsymbol{\theta}_i^{(h')} \sim \mathrm{Dirichlet}\left(\mathbf{n}_i + \boldsymbol{\mu}^{(h)} \tau^{(h)}\right)$ with order restriction, $\boldsymbol{\mu}^{(h)}$ and $\tau^{(h)}$ are the posterior samples from the joint posterior density.

3. Hierarchical multinomial Dirichlet model incorporated uncertainty about order restrictions

3.1 Model specification

We consider adding uncertainty to the model to increase the robustness and flexibility. Let $L_{\text{pos}} = \ell$ be the mode position of cell probabilities. The extension of the hierarchical multinomial Dirichlet model, denoted as M_4, is

$$\mathbf{n}_i \mid \boldsymbol{\theta}_i, L_{\text{pos}} = \ell \overset{\text{ind}}{\sim} \text{Multinomial}\left(\mathbf{n}_{i\cdot}, \boldsymbol{\theta}_i\right), \quad i = 1, \ldots, I, \quad \ell = 1, \ldots, K,$$

$$\boldsymbol{\theta}_i \mid \boldsymbol{\mu}, \tau, L_{\text{pos}} = \ell \overset{\text{ind}}{\sim} \text{Dirichlet}\left(\boldsymbol{\mu}\tau\right), i = 1, \ldots, I, \boldsymbol{\theta}_i \in C_l,$$

$$\pi\left(\boldsymbol{\mu}, \tau \mid L_{\text{pos}} = \ell\right) = \frac{K\left(m_l - 1\right)!\left(K - m_\ell\right)!}{\left(1+\tau\right)^2}, \quad \mu_j > 0, \quad \sum_{j=1}^{K} \mu_j = 1, \quad \boldsymbol{\mu} \in C_{\boldsymbol{\mu}_\ell},$$

where

$$C_\ell = \left\{\boldsymbol{\theta}_i : \theta_{i1} \le \ldots \le \theta_{im_\ell} \ge \ldots \ge \theta_{iK}\right\},$$

$$C_{\boldsymbol{\mu}_\ell} = \left\{\boldsymbol{\mu} : \mu_1 \le \ldots \le \mu_{m_\ell} \ge \ldots \ge \mu_K\right\},$$

and

$$P\left(L_{\text{pos}} = \ell\right) = w_\ell, \quad \sum_{\ell=1}^{K} w_\ell = 1, \quad \ell = 1, \ldots, L.$$

Modes are the same for all areas but we are uncertain about where they are.

Then the joint posterior distribution of $\boldsymbol{\theta}$, $\boldsymbol{\mu}$, and τ, is

$$\begin{aligned}
\pi\left(\boldsymbol{\theta}, \boldsymbol{\mu}, \tau \mid \mathbf{n}\right) &\propto \sum_{L_{\text{pos}}=1}^{L} w_{L_{\text{pos}}} \prod_{i=1}^{I} \left\{ \prod_{j=1}^{K} \theta_{ij}^{n_{ij}} \frac{\prod_{j=1}^{K} \theta_{ij}^{\mu_j \tau - 1} I_{C_{L_{\text{pos}}}} I_{C_{\boldsymbol{\mu}_{L_{\text{pos}}}}}}{D\left(\boldsymbol{\mu}\tau\right) C\left(\boldsymbol{\mu}\tau\right)} \right\} \frac{1}{\left(1+\tau\right)^2} \\
&\propto \sum_{L_{\text{pos}}=1}^{L} w_{L_{\text{pos}}} \prod_{i=1}^{I} \left\{ \frac{\prod_{j=1}^{K} \theta_{ij}^{n_{ij} + \mu_j \tau - 1} I_{C_{L_{\text{pos}}}} I_{C_{\boldsymbol{\mu}_{L_{\text{pos}}}}}}{D\left(\boldsymbol{\mu}\tau\right) C\left(\boldsymbol{\mu}\tau\right)} \right\} \frac{1}{\left(1+\tau\right)^2},
\end{aligned}$$

where $I_{C_{L_{\text{pos}}}}$ and $I_{C_{\boldsymbol{\mu}_{L_{\text{pos}}}}}$ are the indicator functions under that order restriction.

3.2 Estimation of $P\left(L_{\text{pos}} = \ell \mid \mathbf{n}\right)$

To generate samples of $\boldsymbol{\theta}$, $\boldsymbol{\mu}$ and τ, we have to deal with the uncertainty indicator L_{pos}. In M_4, variable L_{pos} has prior $P\left(L_{\text{pos}} = \ell\right) = w_l$ and posterior

$$P\left(L_{\text{pos}} = \ell \mid \mathbf{n}\right) = \frac{w_{L_{\text{pos}}} \int_{\boldsymbol{\mu}} \int_{\tau} \prod_{i=1}^{I} \left\{ \frac{\prod_{j=1}^{K} \theta_{ij}^{n_{ij}+\mu_j \tau - 1} I_{C_{L_{\text{pos}}}} I_{C_{\boldsymbol{\mu}_{L_{\text{pos}}}}}}{D(\boldsymbol{\mu}\tau)\, C(\boldsymbol{\mu}\tau)} \right\} \frac{1}{(1+\tau)^2}\, d\tau d\boldsymbol{\mu}}{\sum_{L_{\text{pos}}=1}^{L} w_{L_{\text{pos}}} \int_{\boldsymbol{\mu}} \int_{\tau} \prod_{i=1}^{I} \left\{ \frac{\prod_{j=1}^{K} \theta_{ij}^{n_{ij}+\mu_j \tau - 1} I_{C_{L_{\text{pos}}}} I_{C_{\boldsymbol{\mu}_{L_{\text{pos}}}}}}{D(\boldsymbol{\mu}\tau)\, C(\boldsymbol{\mu}\tau)} \right\} \frac{1}{(1+\tau)^2}\, d\tau d\boldsymbol{\mu}}.$$

Chen and Nandram (2021) notice the order restrictions will significantly increase the computational difficulty, especially for the marginal likelihood. There is an accuracy-efficiency trade-off. We notice that for each iteration from M1 model, there are two patterns of unimodal structure in $\boldsymbol{\theta}$ for different counties. One is that the normal BMI level has the highest cell probability among five levels, which can be considered as an unimodal structure and the mode is at the second position. Another is that the overweight BMI level has the highest probability, which can be considered as an unimodal structure and the mode is at the third position. We can approximate the posterior sample using information obtained from M1.

Estimation Method:

1. Apply M1 model to the data and acquire posterior samples of θ.
2. For each iteration, count areas whose first cell probabilities are the largest among other cells.
3. In the same iteration, count areas whose second cell probabilities are the largest.
4. Count areas whose third cell probabilities are the largest, until the last cell.
5. Compute the ratio of different cases. For example, we may only have 13 counties whose normal BMI level probabilities are the largest and 22 counties whose overweight probabilities are the largest. Then we have the ratio is 13/22.
6. Compute the average of ratios for overall iterations. Use the average as approximated mixture probabilities.

For example, in our application BMI, 37.2% of $\boldsymbol{\theta}$ has mode at the second position, 62.8% of $\boldsymbol{\theta}$ has mode at the third position. Then we can have $\left(\widehat{L_{\text{pos}} = 2} \mid \mathbf{n}\right) \approx 0.372$ and $P\left(\widehat{L_{\text{pos}} = 3} \mid \mathbf{n}\right) \approx 0.628$ as probabilities to mix samples from M2 (mode at 2nd) and samples from M3 (mode at 3rd) together.

Then

$$\widehat{\text{CPO}}_{i\,(M_4)} \approx \left[\sum_{\ell=1}^{K} P\left(\widehat{L_{\text{pos}} = \ell} \mid n\right) \frac{1}{\widehat{\text{CPO}}_{i\,(L_{\text{pos}}=\ell)}} \right]^{-1},$$

where $\widehat{\text{CPO}}_{i\,(L_{\text{pos}}=\ell)}$ are computed in Section 2.1. In the following numerical example,

$$\widehat{\text{CPO}}_{i\,(M_4)} \approx \left[P\left(\widehat{L_{\text{pos}} = 2} \mid n\right) \frac{1}{\widehat{\text{CPO}}_{i\,(M_2)}} + P\left(\widehat{L_{\text{pos}} = 3} \mid n\right) \frac{1}{\widehat{\text{CPO}}_{i\,(M_3)}} \right]^{-1},$$

without extra computation, taking advantage of known CPOs and the estimated $P\left(L_{\text{pos}} = \ell \mid n\right)$ from the previous section, we can easily acquire the CPO of M_4, as in Appendix A.3.

4. Application to body mass index data

4.1 Body mass index

The performance of our method is studied using the Third National Health and Nutrition Examination Survey, NHANES III. NHANES III is a stratified multistage probability design targeted to obtain a representative sample of the total civilian noninstitutionalized U.S. population age 2 months and older. The sample was selected from households across the United States during the period October 1988 through September 1994. Some individuals area selected with different probabilities. For confidentiality reasons, the final data set for this study uses only the 35 largest counties (from 14 states) with a population of at least 500,000 for selected age categories by sex (male, female) and race (white non-Hispanic, black non-Hispanic, Hispanic, other).

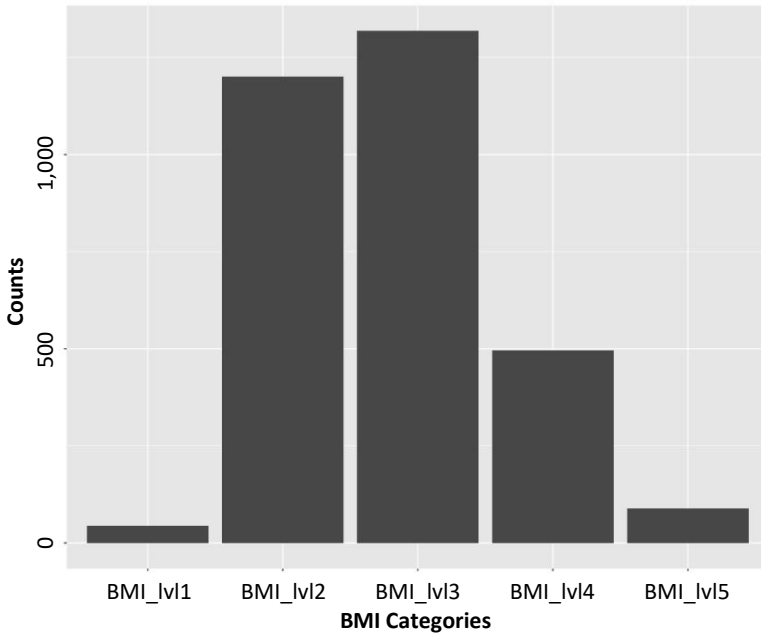
The original sensitive attributes BMI data are transformed to categorical data based on the criteria defined by the Centers for Disease Control (CDC), which are underweight, normal, overweight, obese I, and obese II. If BMI is less than 18.5, it falls within the underweight range. If BMI is 18.5 to $<$ 25, it falls within the normal. If BMI is 25.0 to $<$ 30, it falls within the overweight range. If BMI is 30.0 to $<$ 35, it falls within the obese I range. If BMI is 35.0 or higher, it falls within the obese II range. Our goal is to estimate the proportions of the BMI levels. Table 4.1 gives an illustration of the female BMI data of a few counties, where it can be seen that the cell probability is largest for the normal range and other probabilities roughly tail off on both sides to form the unimodal order restriction. Indeed, there are violations in some counties in the earliest and latest cells.

Thus, for each county, the BMI counts can be assumed to follow a multinomial distribution because each individual person can be assumed to exist independently. Figure 4.1 shows a histogram of all BMI values for females aggregated into a single large sample. It can be clearly seen that the unimodal order restriction holds. Because the data in the individual counties are generally sparse, it is difficult to tell whether the unimodal order restriction holds. However, it is sensible to assume that the same unimodal restriction holds within all the counties. Therefore, we can use multinomial distributions to model the female BMI counts.

Table 4.1
The female BMI in five levels

County ID	BMI_lvl1	BMI_lvl2	BMI_lvl3	BMI_lvl4	BMI_lvl5
1	3	40	37	13	4
2	1	36	38	15	1
3	3	20	49	13	5
⋮	⋮	⋮	⋮	⋮	⋮
35	1	41	41	9	0
Total	45	1,201	1,318	496	89

Figure 4.1 The overall female BMI in five categories.



4.2 Fitting M_1, M_2, M_3 and M_4

4.2.1 MCMC convergence

For each model, we run 20,000 MCMC iterations, take 10,000 as a "burn in" and use every 10th to obtain 1,000 converged posterior samples to maintain consistency. Table 4.2 gives the effective sample sizes of the parameters $\boldsymbol{\mu}$, τ for the model with the order restriction and the general model. The effective sample sizes are almost 1,000. Table 4.3 provides p-values of the Geweke test to check the convergence of the parameters (Geweke, 1991). All p-values are large enough to not reject the null hypothesis that the MCMC is stationary. Then posterior samples can be used for the further inference.

Table 4.2
Effective sample sizes of μ and τ

	μ_1	μ_2	μ_3	μ_4	μ_5	τ
M_1	1,000	1,123.7	1,000	1,000	895.4	1,000
M_2 (Mode at 2nd)	1,000	1,000	1,000	1,000	1,150.2	1,000
M_3 (Mode at 3rd)	1,000	887	889	1,000	1,173.9	1,000

Table 4.3
P values of Geweke test for μ and τ

	μ_1	μ_2	μ_3	μ_4	μ_5	τ
M_1	0.623	0.558	0.899	0.767	0.959	0.514
M_2 (Mode at 2nd)	0.964	0.705	0.507	0.511	0.837	0.999
M_3 (Mode at 3rd)	0.817	0.559	0.580	0.557	0.812	0.516

4.2.2 Model comparison

With the approximate mixture probabilities, we mix posterior samples of M_2 and M_3 together to construct samples of M_4.

We provide posterior mean (PM), posterior standard deviation (PSD) and coefficient of variation (CV) of $\boldsymbol{\theta}$'s for all counties, which can be found in Appendix A.4.

To compare model difference visually, we present the posterior densities plots about different counties in those models as Figure 4.2, Figure 4.3, Figure 4.4 and Figure 4.5. We use different colors to indicate five BMI levels and dashed lines for the posterior means. Due to different capability of borrowing information among areas, we can see different flatness of posterior density curves in the models. With different order restriction assumptions, those posterior density curves center at different places and may overlap differently. We mainly focus on density curves of normal BMI and overweight BMI, since the modal position might be second or third.

In Figure 4.2 has posterior density plots for County 2 applying different models. The number of observations with normal BMI level, which is 36, is close to the number of observations with overweight BMI level, which is 38. The unimodal order restriction may not hold in County 2. Maybe for this reason, there is a significant overlap between normal level and overweight level in the first plot after applying M_1 to our BMI data. The second plot and the third plot show much less overlap in density curves, due to the strong order restriction assumption. The last plot, which is the density curve from M_4, is similar to the density curves in M_3. Based on the observations in County 2, the order restriction that the modal position is at the third may be reasonable. The density curves in M_3 and M_4 may be appropriate for County 2.

In Figure 4.3 has posterior density plots for County 3 applying different models. Unlike in County 2, the density curves of θ from model M_1 in County 3 shows a very strong unimodality because we have 49 people in overweight BMI level which dominates this county. The second plot from M_2, which assumes that the mode is at normal BMI level, has a significant overlap. Its order restriction assumption that the modal position is at the second position may not hold in this county. The third plot from M_3, which assumes that the mode is at overweight BMI level, is similar as the density curve in M_1. The posterior mean of normal BMI level probability is higher than in M_1. This phenomenon can be considered as an evidence that M_3 has a stronger borrowing ability than M_1. Overall, the modal position among 35 counties may be at the third. M_3 can borrow more information among those counties than other models. Then the last plot, which is the density curve from M_4, has a little overlap. But the unimodal pattern is still in M_4.

In Figure 4.4, they are posterior density plots for County 13 applying different models. Only M_2 with an assumption that the mode is at normal BMI level does not show a significant overlap. Since more people are at overweight BMI level, that assumption may be validate in County 13.

Figure 4.5 provides posterior density plots for County 35, which has almost same amount of people in normal and overweight BMI level. M_2 and M_3 with different unimodal assumptions have opposite conclusion about normal and overweight probabilities. In this county, M_1 and M_4 may be better models.

Figure 4.2 Posterior densities of θ for county 2 showing different order restrictions under different models.

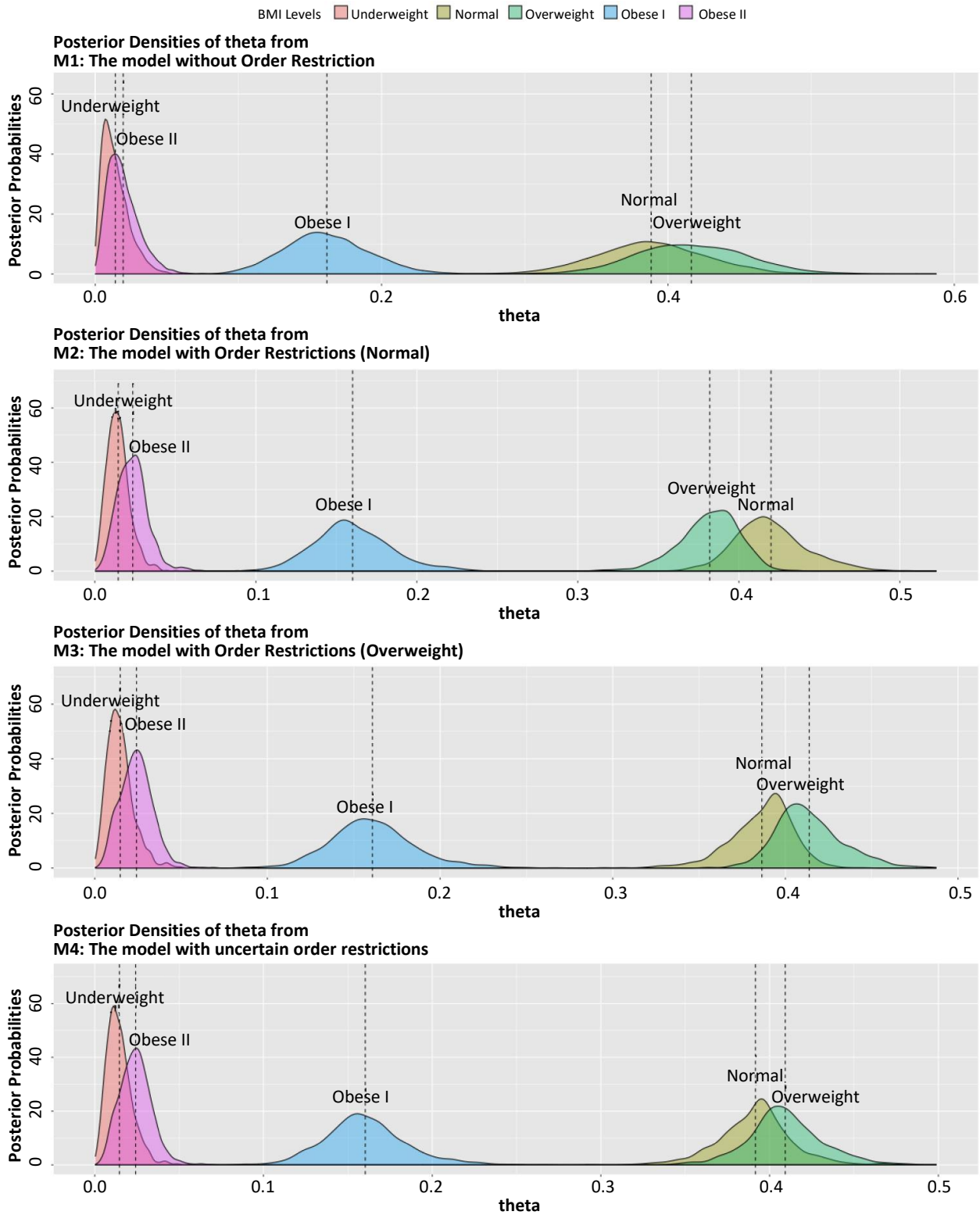


Figure 4.3 Posterior densities of θ for county 3 showing different order restrictions under different models.

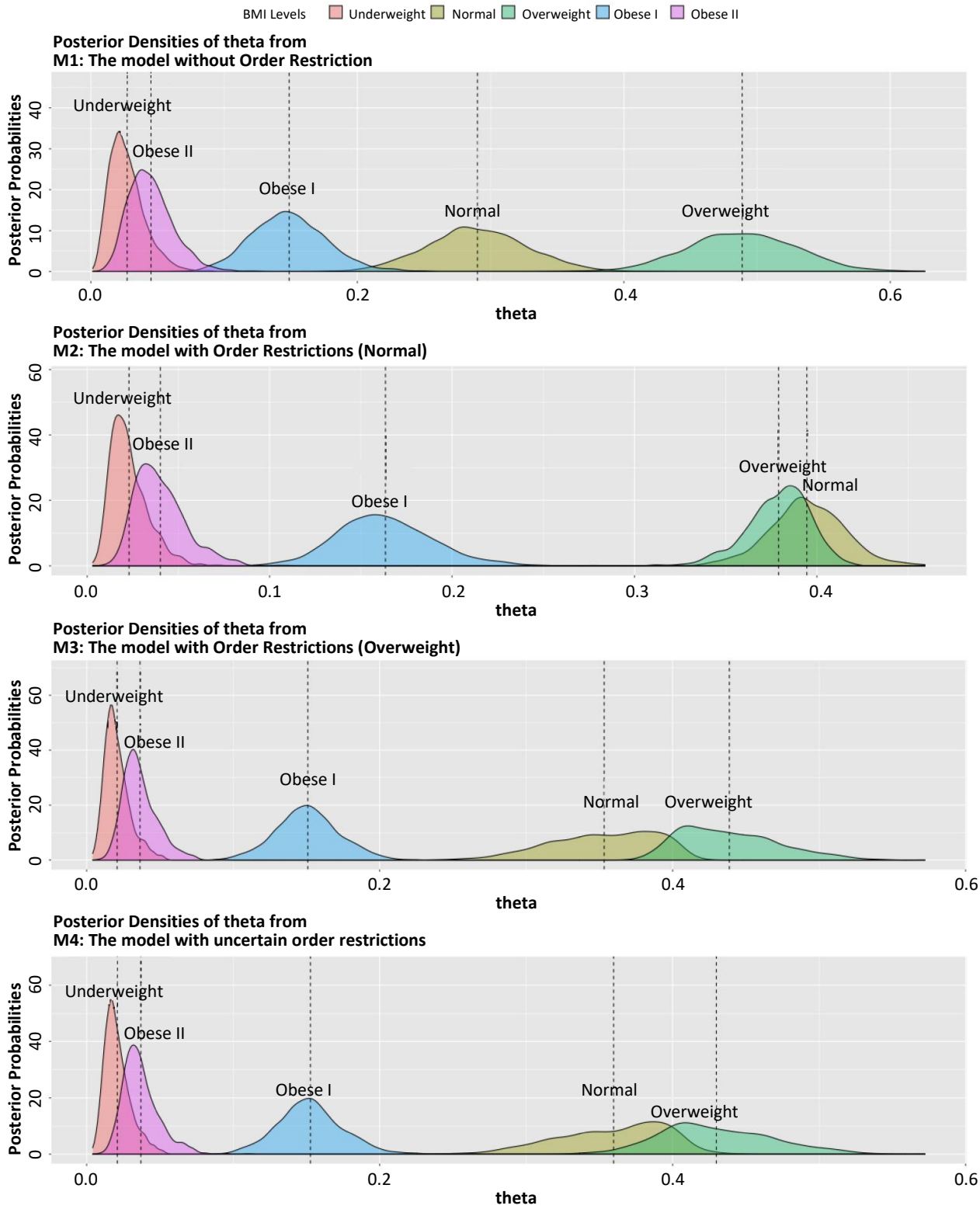


Figure 4.4 Posterior densities of θ for county 13 showing different order restrictions under different models.

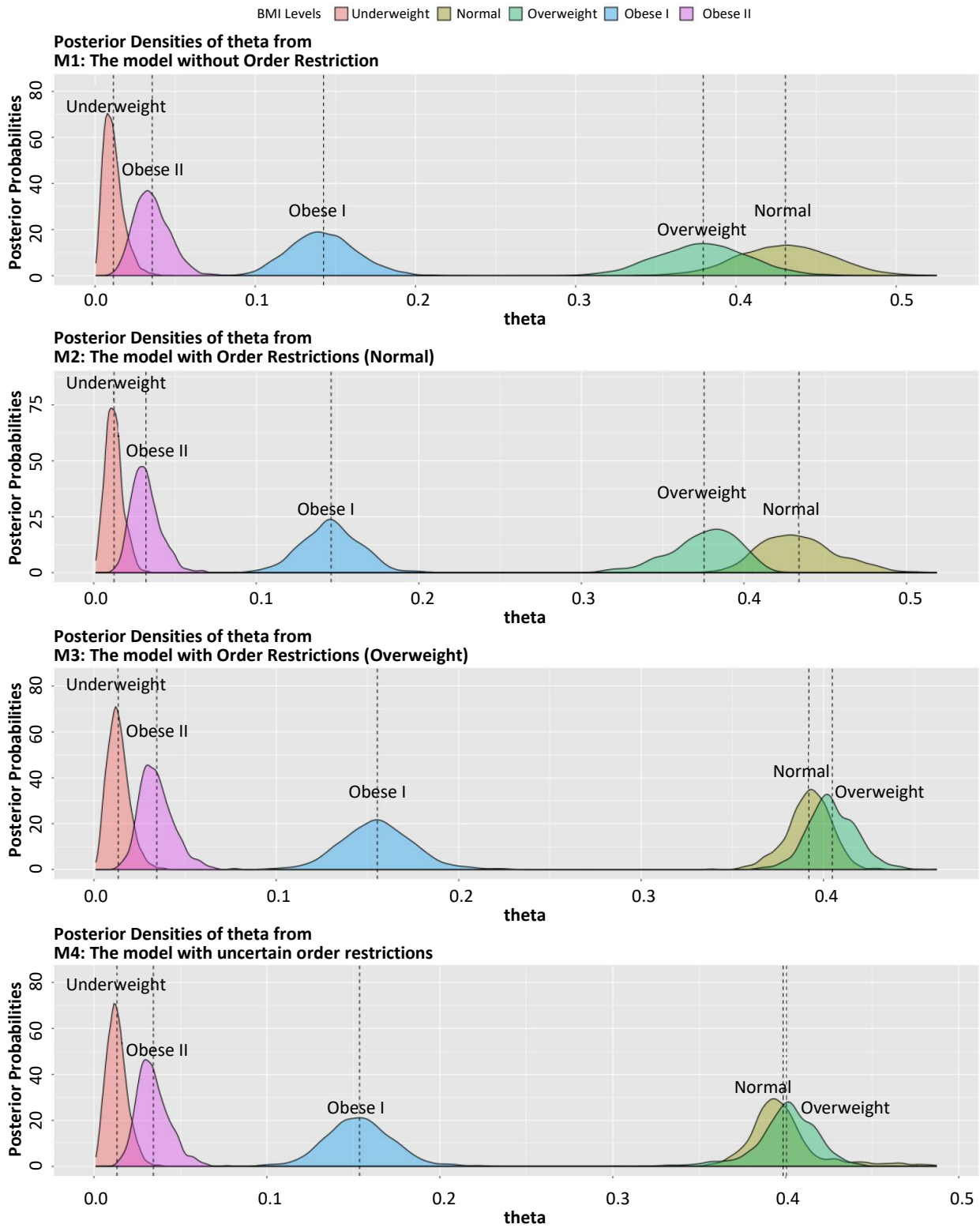
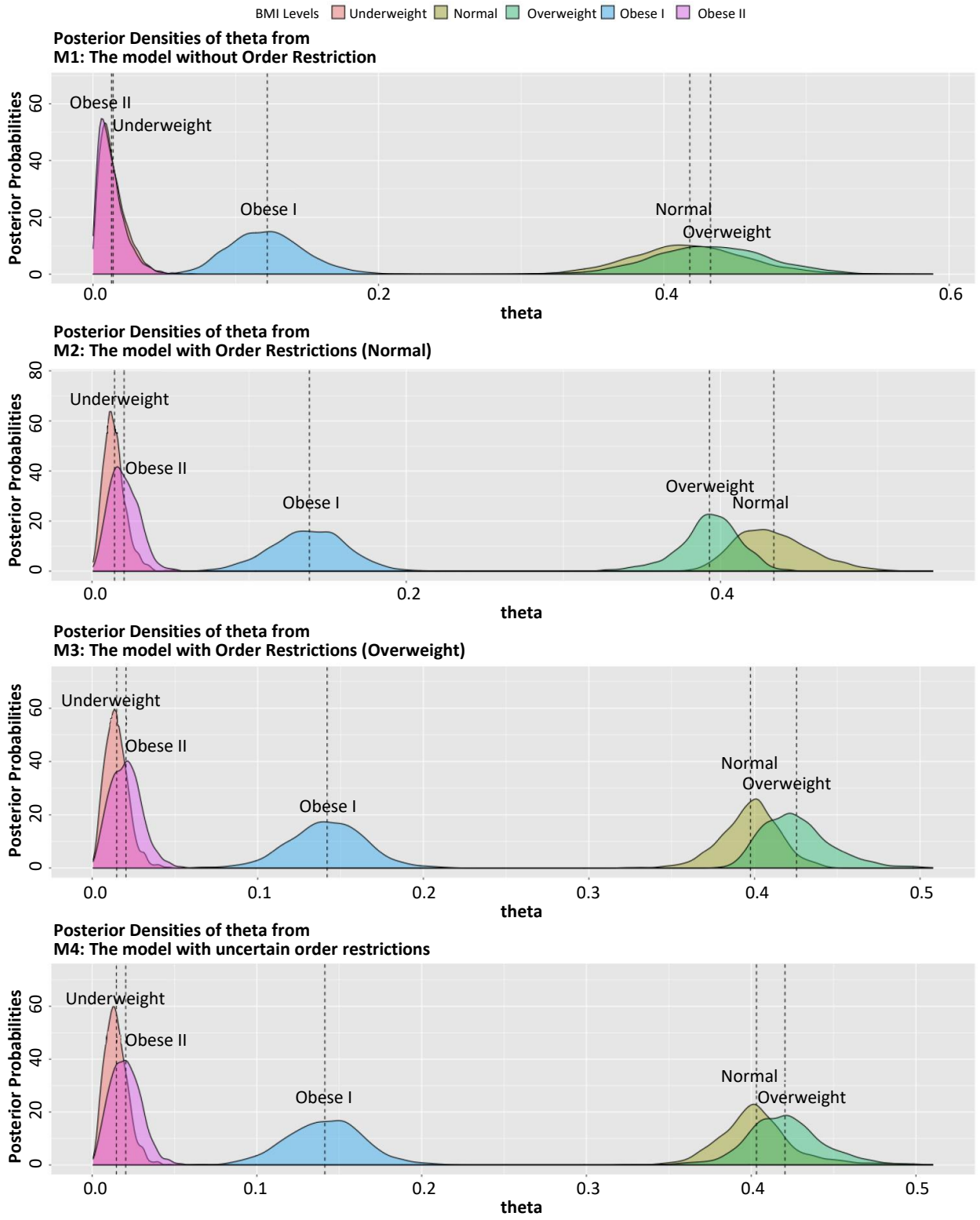


Figure 4.5 Posterior densities of θ for county 35 showing different order restrictions under different models.



Overall, the model with order restrictions, M_2 and M_3, can borrow more information among areas than the model without order restriction, M_1. The model with uncertain order restriction, M_4, borrow less information among areas than M_2 or M_3. For this reason, M_2 and M_3 have sharper posterior density curves than M_1, M_4 has slightly flatter posterior density curves than M_2 and M_3. For the same reason, as shown in Table 4.4, M_1 has the largest total variance, which is the sum of posterior variance of all counties' cell probabilities. M_2 and M_3 have the smallest variance due to its strong unimodal order restriction assumption. M_4's variance is between M_1 and M_3 (or M_2) since M_4 is a mixture of M_2 and M_3.

Table 4.4
Total variance of θ

M_1	M_2 (mode at normal)	M_3 (mode at overweight)	M_4
0.172	0.063	0.069	0.107

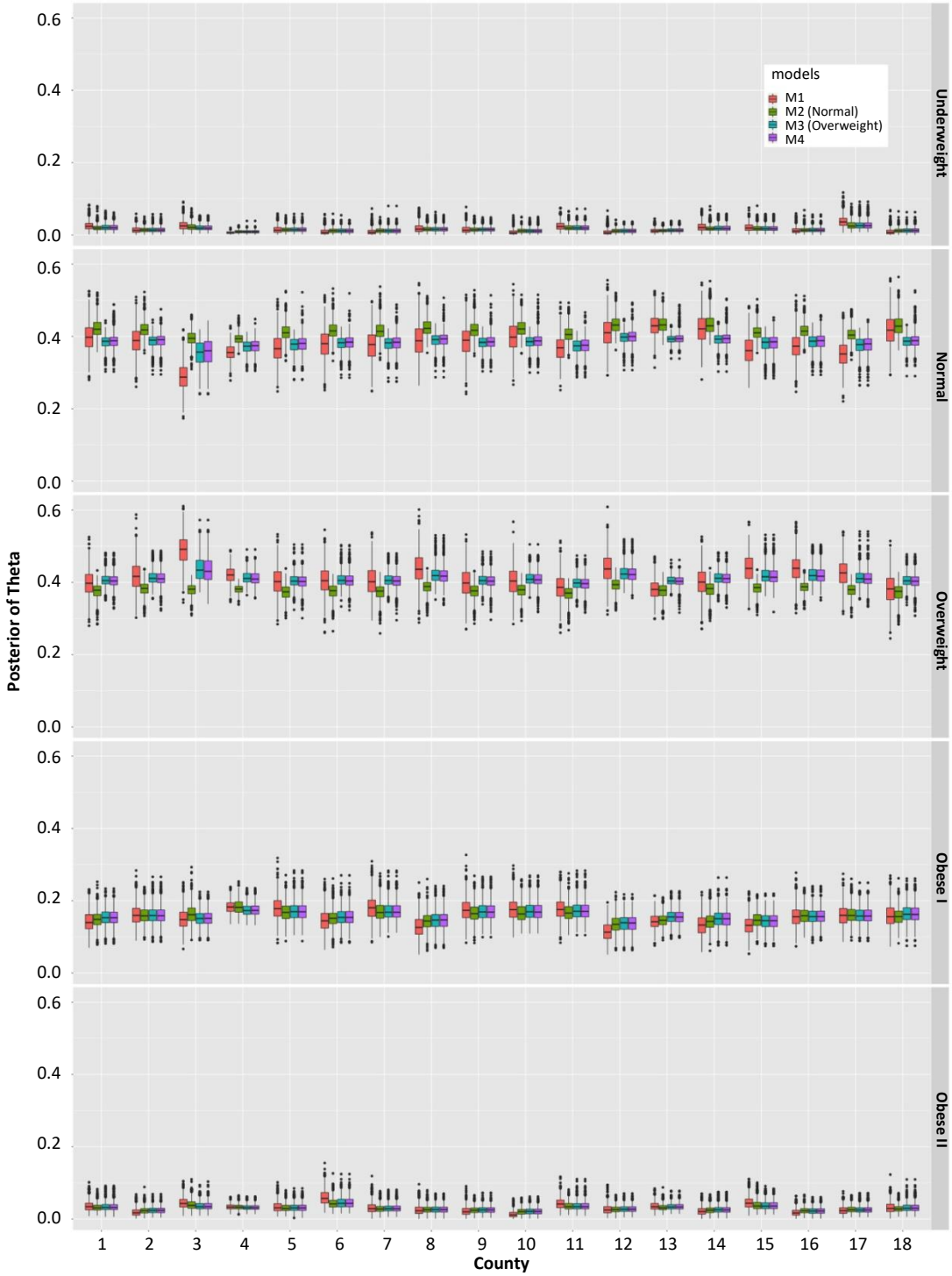
Figure 4.6 and Figure 4.7 are boxplots of θ's posterior samples. The first (Underweight) and last (Obese II) blocks show that different models do not have much difference in estimating the cell probabilities of underweight, normal, and obese I. In the box plots, short line segments from M_2, M_3, and M_4 and long line segments from M_1 show that the models with order restrictions (M_2, M_3, M_4) have smaller variances than the model without order restriction (M_1). The models with order restrictions can borrow more information than the model without order restriction. The differences between each box of M_1 are larger than the differences in M_2, M_3, and M_4. In other word, the differences between posterior mean of each county in M_1 are larger than other models'. It proves that the models with order restrictions borrow more information among areas than the model without order restriction.

In Figure 4.8, we have some regression lines to show the overall posterior standard deviation comparison among those models. The black dashed line is a reference line whose slope is one. The first plot shows a comparison between M_1 and M_3 (mode at overweight). All of regression lines are above the reference line, which means that M_3 (mode at overweight) has smaller standard deviation. We gain higher precision on estimation of cell probabilities among 35 counties in M_3. The second plot shows a comparison between M_2 (mode at normal) and M_3 (mode at overweight). The regression lines about underweight, Obese I and Obese II are around the reference line. Only the regression line about overweight shows significant difference. It means M_3 (mode at overweight) is slightly better than M_2 (mode at normal). In other word, the assumption that overweight BMI probability is the highest may be more reasonable. The last two plots in Figure 4.8 is a comparison between M_2 (mode at normal) and M_4, M_3 (mode at overweight) and M_4. M_4's performance is slightly worse than M_3 and M_2.

In Figure 4.9, we use different symbols to represent different models' CPO for all 35 counties. In BMI data, County 4 has the largest population, which shows lowest CPO value among others. It is known that low CPO values suggest possible outliers, high-leverage and influential observations. Due to the borrowing feature from the models, County 3 has a low CPO which may be affected by County 4. For

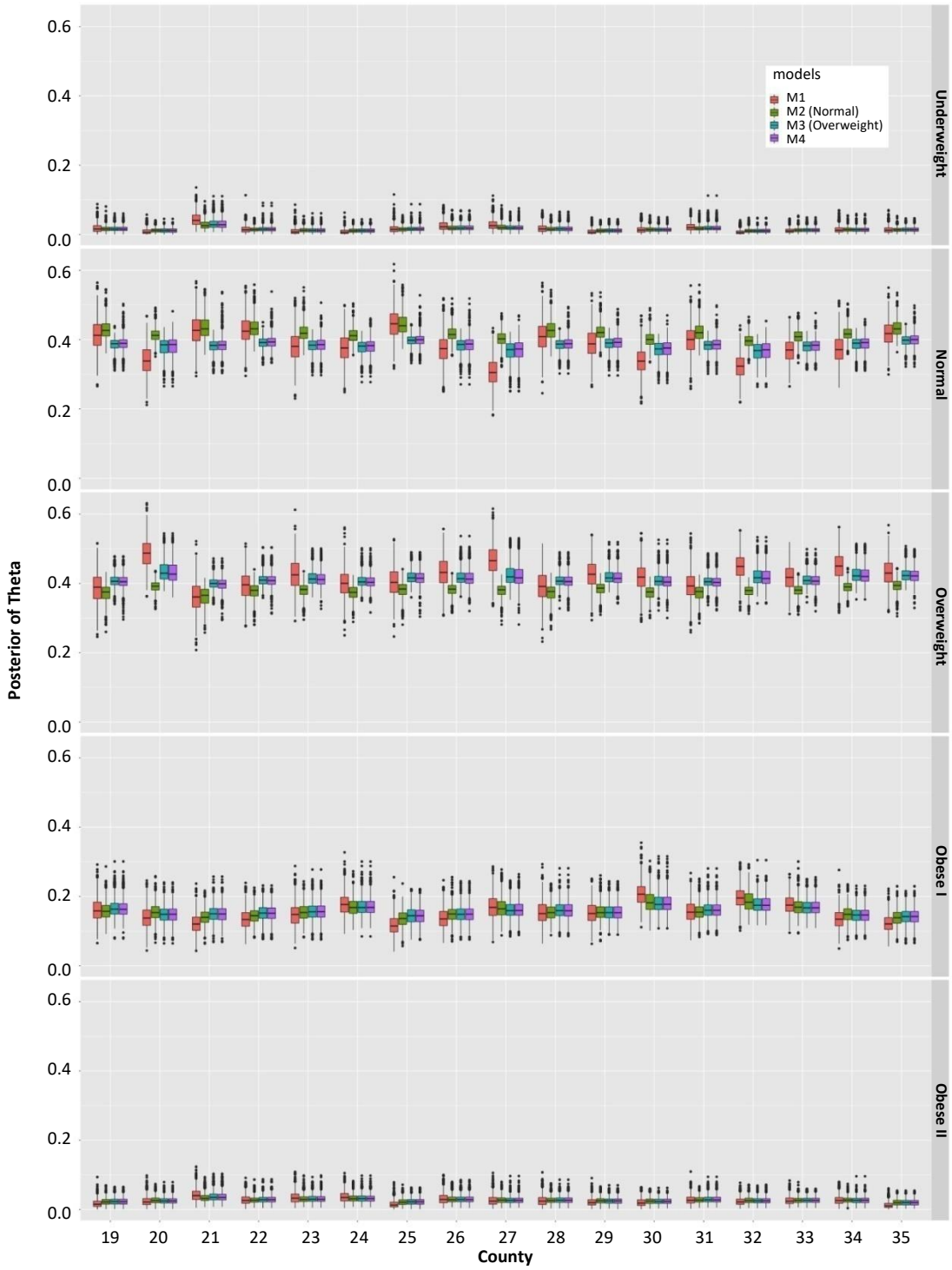
most counties, the model with order restriction which assumes the mode is at overweight position can have large CPO, compared with other models. As a summary, in Table 4.5, M_2 (mode at overweight) has the largest LMPL, which should the "best" model for our BMI data.

Figure 4.6 Posterior of θ : Part I.



This is a county-wise comparison for different BMI categories under different models.

Figure 4.7 Posterior of θ : Part II.



This is a county-wise comparison for different BMI categories under different models.

Figure 4.8 Standard deviation comparison between those models to show improvement.

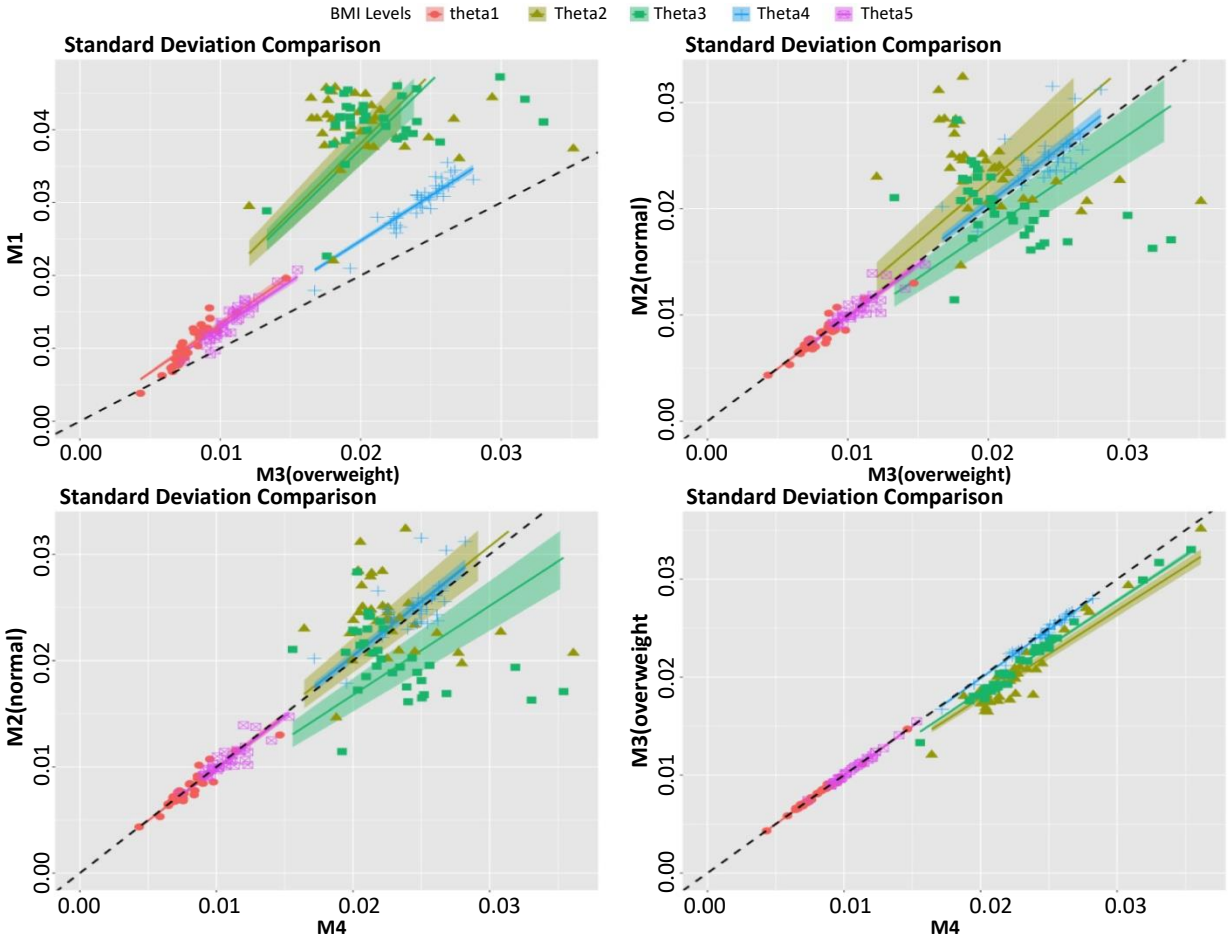
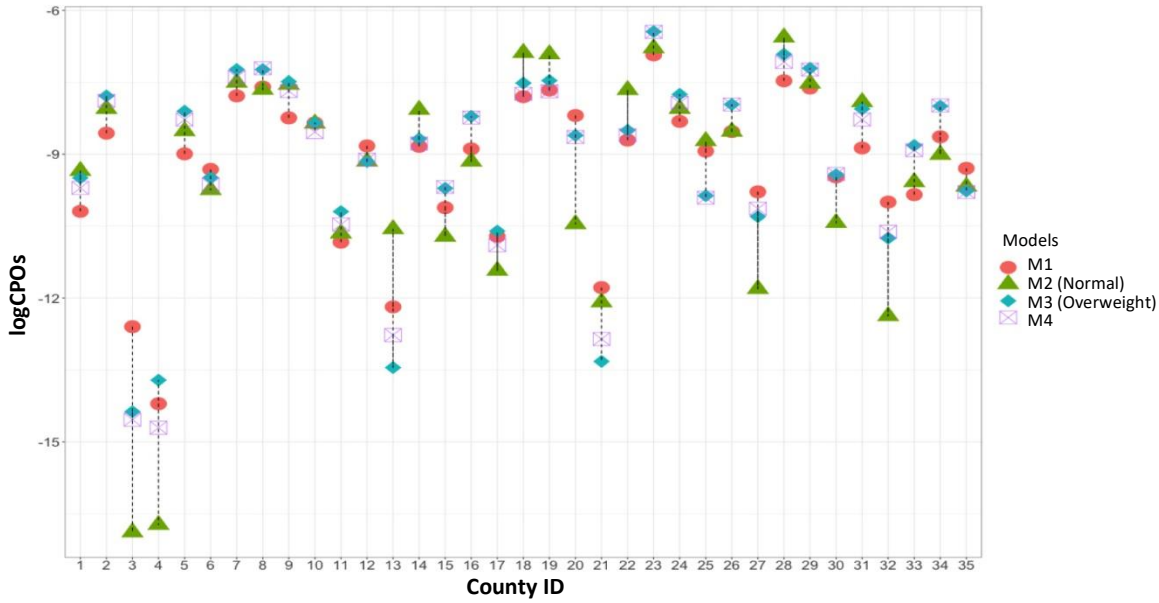


Figure 4.9 CPOs for 35 Counties under different models.



Note: Lower CPO suggests possible outliers, high-leverage and influential observations.

Table 4.5
LPML, comparison of the four models using LPML

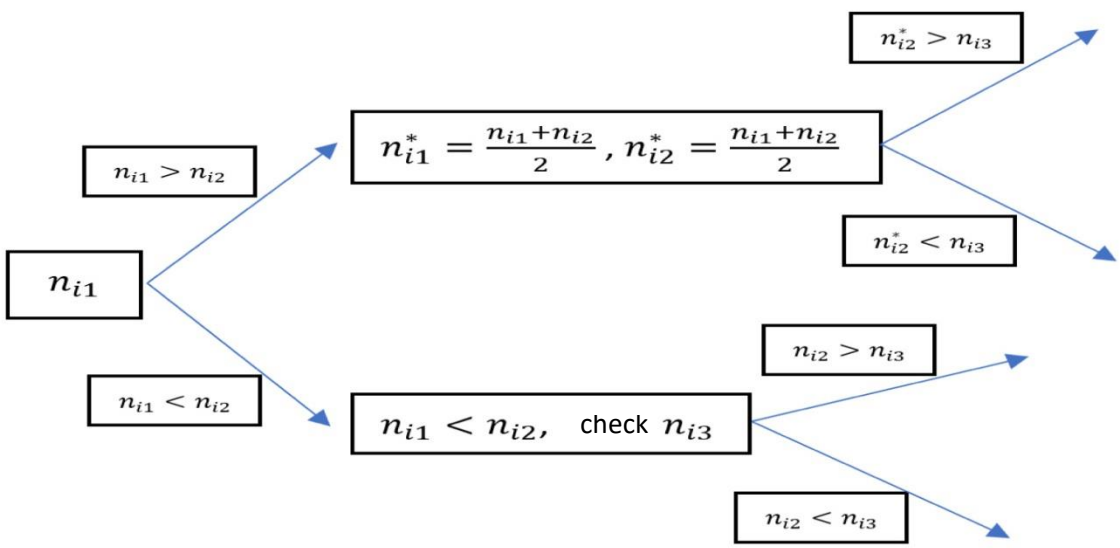
M_1	M_2 (mode at normal)	M_3 (mode at overweight)	M_4
-326.76	-331.88	-318.26	-329.58

5. Simulated BMI

To have a better comparison between those models, Chen and Nandram (2021) construct a simulated data transformed from BMI using the idea of Pool-Adjacent-Violators Algorithm (PAVA) to have strong order restrictions as $\theta_1 \leq \ldots \leq \theta_m \geq \ldots \geq \theta_K$, (Mair, Hornik and de Leeuw, 2009). It is a simple iterative algorithm for solving the quadratic problem.

Generally, given a sequence of n data points $y_1, \ldots, y_n$, we start with y_1 on the left. We move to the right until we encounter the first violation $y_i > y_{i+1}$. Then we replace this pair by their average, and back-average to the left as needed, to get monotonicity. We continue this process to the right, until finally we reach y_n. We can have a reconstructed data set to fit our order restrictions better. Fitting models to the simulated data, we can discover the advantage of hierarchical multinomial-Dirichlet model with order restrictions easily.

Figure 5.1 Simulation method to have the unimodal order restriction.



Here, for each county, we start from BMI level 1 to the mode using PAVA to create an increasing sequence. Then from the mode to BMI level 5, we apply PAVA to create a decreasing sequence. To make sure that each BMI level has an integer number, we take the nearest integer that is larger than the mode to replace the mode, and take the nearest integer that is smaller than n_{ij} (except the mode) to replace those

non-modes. Now our assembled BMI data have strong order restrictions. But we also notice that our current approach cannot be used for a general case to create an unimodal structure. It works for BMI data when the numbers of level 2 and level 3 are significantly larger than others. Now we have a simulated BMI data which mode is at the third position (overweight).

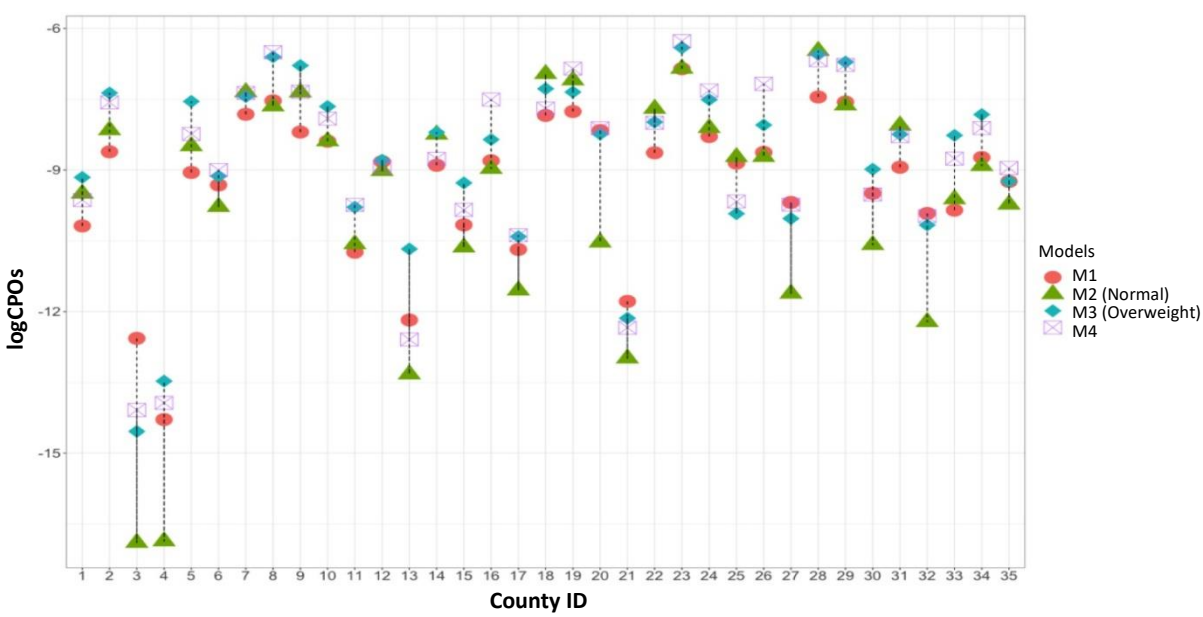
Table 5.1
LPMLs of model M_1, M_2, M_3, and M_4 for simulated data, comparison of the four models using LPML

M_1	M_2 (mode at normal)	M_3 (mode at overweight)	M_4
-319.83	-330.73	-310.39	-311.26

Since the mode is at the third position, the LPML of M_3 is significantly larger than others, which is -310.39. The LPML of M_4 is -311.26, due to the robustness of M_4. The LPML of M_2 is the smallest, which is -330.73. The LPML of M_1 is -319.83. The LPMLs show that the model with order restrictions can have the best performance if the unimodal assumption is correct. Model M_4, which incorporates uncertainty about order, has a similar performance as Model M_3. In Figure 5.2, M_3 and M_4 have consistently large CPO values for 35 counties among those models. M_2 have lowest CPO values at County 3 and 4, which suggests possible outliers, high-leverage and influential observations. For most of counties, M_3 has the largest CPOs and M_2 has the smallest CPOs because of the order restriction assumption may be correct in M_3 but not in M_2.

In the simulated BMI data, CPO and LPML are proved to be able to select more adequate models. Model M_4 is robust and consistent for most cases.

Figure 5.2 CPOs for 35 counties under different models (simulation).



Note: Lower CPO suggests possible outliers, high-leverage and influential observations.

6. Concluding remarks

The Dirichlet multinomial model with mixed order restrictions is an extension of M_2. It increases the robustness and flexibility due to its uncertainty. We have also shown how to acquire samples of the model with mixed order restriction. In our application and simulation, we find that, with the uncertainty, the Dirichlet multinomial model with mixed order restrictions may be the best model for all cases with varied unknown unimodality. For most cases, we could not know the unimodal order restriction, even if we believe it exists. Bringing uncertainty to the model is necessary. We also notice that due to its complexity, it is hard to compute its marginal likelihood. We show a method to estimate the posterior probabilities of the mode location, which is $P\left(L_{\text{pos}} = \ell \mid \mathbf{n}\right)$. But there is a precision-efficiency tradeoff.

However, as shown in Figure 4.2 and Figure 4.3, the same unimodal order restriction for all counties may be still strong even with uncertainty. Some counties have more people in the normal BMI level, and some counties have more people in the overweight BMI level. Nandram and Sedransk (1995) and Nandram, Sedransk and Smith (1997) presented a good discussion about unimodal order restriction in a stratified population. With the help of uncertainty, they made inference about the proportion of firms and fish belonging to each of several classes when there are unimodal order relations among the proportions. In that paper, the hyperparameters are specified and they did not have a small area estimation problem; our problem is much more difficult even we consider a similar uncertainty model structure.

In Section 4.2.2, the model with fixed order restrictions is a better model for BMI data because of its largest LPML. But without any background, assuming the modal position is risky and may cause the wrong inference. The multinomial Dirichlet model with order restrictions, incorporating uncertainty, can reduce the risk and is more robust. In the simulation, Model M_2 is the best model for the simulated BMI data. Model M_4 shows a better consistency for the simulated BMI data and the real BMI data.

The final BMI data set for this study uses only the 35 largest counties with a population of at least 500,000 for selected age categories by sex (male, female) and race (white non-Hispanic, black non-Hispanic, Hispanic, other). We can easily apply our method to the small domains formed by on race, age and sex, such as the male-Hispanic BMI data. But the cells of the multinomial tables will become sparse. We can eliminate some counties that become small or we can combine some counties. However, due to the structures of multinomial-Dirichlet models with order restrictions, we cannot add race, age and sex as covariates into the model.

Since the BMI data are from the survey sampling and individuals are selected with different probabilities, we should not ignore the survey weights. It is possible to incorporate the survey weights into our model as well. Let W_{ig} denote the survey weights, adding up to the population size within each county, $i = 1, \ldots, \ell$, sample index $g = 1, \ldots, n_i$ and cell index $j = 1, \ldots, K$. Yang (2021) provided adjusted weights are

$$\omega_{ig} = n_i \frac{W_{ig}}{\sum_{g=1}^{n_i} W_{ig}},$$

and $\sum_{g=1}^{n_i} \omega_{ig} = n_{i.} = \sum_{j=1}^{K} n_{ij}$. Yang (2021) used weighted likelihood distributions for a single multinomial model, see also Nandram, Choi and Liu (2021). Yang (2021) found out there is a very small difference between normalized and unnormalized weighed likelihood.

We can transform BMI data using the adjusted weights into adjusted counts. Let I_{igj} be the BMI category indicator for individual g in county $i, i = 1, \ldots, \ell$ at cell $j, j = 1, \ldots, K$. We define $I_{igj} = 0$ or 1 with $\sum_{j=1}^{K} I_{igj} = 1$, for example, if a person responds in cell j, a one is scored and all other cells have zeros. For simplification, we can have the unnormalized weighted joint posterior distribution as

$$\pi\left(\boldsymbol{\theta}, \boldsymbol{\mu}, \tau, \mathbf{p}, \phi \mid \mathbf{n}\right) \propto \prod_{i=1}^{\ell} \left\{ \frac{\left(\sum_{j=1}^{K} \sum_{g=1}^{n_i} I_{igj} \omega_{ig}\right)!}{\prod_{j=1}^{K} \left(\sum_{g=1}^{n_i} I_{igj} \omega_{ig}\right)!} \prod_{j=1}^{K} \theta_{ij}^{\sum_{g=1}^{n_i} I_{igj} w_{ig}} \left[p_i \frac{\text{Dirichlet}\left(\boldsymbol{\mu}\tau\right)}{\int_{\theta_i \in C} \text{Dirichlet}\left(\boldsymbol{\mu}\tau\right) d\theta_i} \frac{\left(K-1\right)!}{\left(1+\tau\right)^2} + \left(1-p_i\right) \text{Dirichlet}\left(1, \ldots, 1\right) \right] \frac{p_i^{\phi\tau_0 - 1} \left(1-p_i\right)^{\left(1-\phi\right)\tau_0 - 1}}{B\left(\phi\tau_0, \left(1-\phi\right)\tau_0\right)} \right\}.$$

Our approaches can be applied to the adjusted counts directly.

It is possible to relax the unimodal order restriction somewhat. One can restrict the position of the mode without any ordering on its left or right, we can still have the mode at 2 or 3 for the BMI data to provide a model with uncertainty about the modal position. This can be done in the same spirit as in our current work.

We notice the same unimodal structure across all counties is not satisfied. Borrowing information across those areas may have a negative effect to model inference. Neuenschwander, Wandel, Roychoudhury and Bailey (2016) presented a different approach to increase the model robustness in drug development. They proposed the exchangeability nonexchangeability (EXNEX) approach to reduce the risk of too much shrinkage and excessive borrowing for extreme strata. We can borrow their approach to increase our model robustness. But we believe it is very difficult to make inference using the Dirichlet multinomial model with EXNEX prior because the model complexity increases significantly.

Appendix

A.1 Gibbs sampler for $\boldsymbol{\mu}$ and τ in M_2 and M_3

We present griddy Gibbs sampler, a Markov chain Monte Carlo (MCMC) algorithm, for $\boldsymbol{\mu}$ with the order restriction and τ.

Liu and Sabatti (2000) presented a comprehensive discussion of the general Gibbs sampler which is more efficient Markov chain Monte Carlo method for Bayesian inference. They explored its connection with the multigrid Monte Carlo method and its use in designing more efficient samplers. Gibbs sampler may be more efficient in our hierarchical model. Therefore we use Gibbs sampler to generate the posterior samples for the Bayesian inference.

We present the modified Gibbs sampler for $\boldsymbol{\mu} \in C_{\boldsymbol{\mu}}$ and τ. The joint posterior density is

$$\pi\left(\boldsymbol{\theta}, \boldsymbol{\mu}, \tau \mid \mathbf{n}\right) \propto \prod_{i=1}^{I} \left\{ \frac{\prod_{j=1}^{K} \theta_{ij}^{n_{ij}+\mu_j\tau-1} I_C I_{C_{\boldsymbol{\mu}}}}{D\left(\boldsymbol{\mu}\tau\right) C\left(\boldsymbol{\mu}\tau\right)} \right\} \frac{1}{(1+\tau)^2},$$

where

$$C\left(\boldsymbol{\mu}\tau\right) = \int_{\boldsymbol{\theta}_i \in C} \frac{\Gamma\left(\sum_{j=1}^{K} \mu_j \tau\right)}{\prod_{j=1}^{K} \Gamma\left(\mu_j \tau\right)} \prod_{j=1}^{K} \theta_{ij}^{\mu_j\tau-1} d\boldsymbol{\theta}_i.$$

There is no recognizable conditional distribution of $\boldsymbol{\mu}$ and τ to generate samples. So we use grid method to draw $\boldsymbol{\mu}$ and τ from $\pi\left(\boldsymbol{\mu}, \tau \mid \mathbf{n}\right)$ after integrating with respect to $\boldsymbol{\theta}$, we get

$$\begin{aligned} \pi\left(\boldsymbol{\mu}, \tau \mid \mathbf{n}\right) &\propto \prod_{i=1}^{I} \left\{ \frac{D\left(\boldsymbol{\mu}\tau + \mathbf{n}_i\right) C\left(\boldsymbol{\mu}\tau + \mathbf{n}_i\right)}{D\left(\boldsymbol{\mu}\tau\right) C\left(\boldsymbol{\mu}\tau\right)} \right\} \frac{I_{C_{\boldsymbol{\mu}}}}{(1+\tau)^2} \\ &\propto \prod_{i=1}^{I} \left\{ \frac{\int_{\boldsymbol{\theta}_i \in C} \prod_{j=1}^{K} \theta_{ij}^{\mu_j\tau+n_{ij}-1} d\boldsymbol{\theta}_i}{\int_{\boldsymbol{\theta}_i \in C} \prod_{j=1}^{K} \theta_{ij}^{\mu_j\tau-1} d\boldsymbol{\theta}_i} \right\} \frac{I_{C_{\boldsymbol{\mu}}}}{(1+\tau)^2}. \end{aligned}$$

Chen and Shao (1997) mentioned that importance sampling could be used to estimate the ratio,

$$\frac{\int_{\boldsymbol{\theta}_i \in C} \prod_{j=1}^{K} \theta_{ij}^{\mu_j\tau+n_{ij}-1} d\boldsymbol{\theta}_i}{\int_{\boldsymbol{\theta}_i \in C} \prod_{j=1}^{K} \theta_{ij}^{\mu_j\tau-1} d\boldsymbol{\theta}_i}.$$

We consider Dirichlet$\left(r\bar{n}_j\right)$ as our importance of all counties function, where r is an adjustable ratio and

$$\bar{n}_j = \frac{\sum_{i=1}^{I} n_{ij}}{I}.$$

It combines information together. Since our importance function does not depend on the unknown $\boldsymbol{\mu}$ and τ, we can generate one set of numbers for all iterations. In our numerical example, it has been proved as an efficient way to generate posterior samples.

Gibbs sampler steps:

1. Draw τ from $\pi\left(\tau \mid \boldsymbol{\mu}, \mathbf{n}\right)$;
2. For j from $m-1$ to 1, draw μ_j from $\pi\left(\mu_j \mid \boldsymbol{\mu}^{(-j)}, \tau, \mathbf{n}\right)$, where
$$0 < \mu_j < \min\left\{ \mu_{j+1}, \frac{1-\sum_{t=1,\, t\neq m,\, t\neq j}^{K} \mu_t}{2} \right\};$$
3. For j from $m+1$ to K, draw μ_j from $\pi\left(\mu_j \mid \boldsymbol{\mu}^{(-j)}, \tau, \mathbf{n}\right)$, where
$$0 < \mu_j < \min\left\{ \mu_{j-1}, \frac{1-\sum_{t=1,\, t\neq m,\, t\neq j}^{K} \mu_t}{2} \right\};$$
4. Get $\mu_m = 1 - \sum_{j=1,\, j\neq m}^{K} \mu_j$, repeat Step 1 to Step 4 until convergence,
$$\mu^{(-j)} = \left(\mu_1, \ldots, \mu_{j-1}, \mu_{j+1}, \ldots, \mu_K\right).$$

A.2 Sampling $\boldsymbol{\theta}$ in M_2 and M_3

The posterior of $\boldsymbol{\theta}$ has a recognizable distribution, which is the Dirichlet distribution with the order restriction. Instead of drawing samples directly from the Dirichlet distribution with the order restriction, Chen and Nandram (2019) present a direct sampling from truncated Gamma distributions, where Nadarajah and Kotz (2006) offered a method for truncated Gamma.

Denote $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)$, if $0 \le \theta_1 \le \theta_2 \le \ldots \le \theta_m \ge \ldots \ge \theta_K$ and the mode is θ_m, then we assume $0 \le \beta_1 \le \beta_2 \le \ldots \le \beta_m \ge \ldots \ge \beta_K$, the mode is β_m.

Steps of sampling θ from Dirichlet$(\alpha_1, \ldots, \alpha_K)$:

1. Draw $\beta_m \sim \text{Gamma}(\alpha_m, 1)$, where $0 \le \beta_m < \infty$;
2. Draw from β_{m-1} to β_1,
   $\beta_{m-1} \sim$ Truncated Gamma$(\alpha_{m-1}, 1)$, where $0 \le \beta_{m-1} \le \beta_m$,
   …
   $\beta_1 \sim$ Truncated Gamma$(\alpha_1, 1)$, where $0 \le \beta_1 \le \beta_2$;
3. Draw from β_{m+1} to β_K,
   $\beta_{m+1} \sim$ Truncated Gamma$(\alpha_{m+1}, 1)$, where $0 \le \beta_{m+1} \le \beta_m$,
   …
   $\beta_K \sim$ Truncated Gamma$(\alpha_K, 1)$, where $0 \le \beta_K \le \beta_{K-1}$.

   Then,

$$\theta_1 = \frac{\beta_1}{\beta_1 + \beta_2 + \ldots + \beta_K}, \ldots, \theta_{K-1} = \frac{\beta_{K-1}}{\beta_1 + \beta_2 + \ldots + \beta_K}, \theta_K = 1 - \sum_{i=1}^{K-1} \theta_i.$$

A.3 Bayesian diagnostics of M_2, M_3, and M_4

Since the only difference between M_2 and M_3 is the order restriction assumption and the CPOs of M_2 and M_3 are similar, we only present the CPO of M_2 here,

$$\begin{aligned}
\widehat{\text{CPO}}_{i\,(M_2)} &= \left[\frac{1}{M} \sum_{h=1}^{M} \frac{\prod_{j=1}^{K} n_{ij}!}{n_{i.}!} \frac{D\left(\boldsymbol{\mu}^{(h)}\tau^{(h)}\right) C\left(\boldsymbol{\mu}^{(h)}\tau^{(h)}\right)}{D\left(\mathbf{n}_i + \boldsymbol{\mu}^{(h)}\tau^{(h)}\right) C\left(\mathbf{n}_i + \boldsymbol{\mu}^{(h)}\tau^{(h)}\right)} \right]^{-1} \\
&= \left[\frac{1}{M} \sum_{h=1}^{M} \frac{\prod_{j=1}^{K} n_{ij}!}{n_{i.}!} \frac{\int_{\theta_i \in C} \prod_{j=1}^{K} \theta_{ij}^{\mu_{(h)}\tau_{(h)}-1} d\,\boldsymbol{\theta}_i}{\int_{\theta_i \in C} \prod_{j=1}^{K} \theta_{ij}^{n_{ij}+\mu_{(h)}\tau_{(h)}-1} d\,\boldsymbol{\theta}_i} \right]^{-1} \\
&= \left[\frac{1}{M} \sum_{h=1}^{M} \frac{\prod_{j=1}^{K} n_{ij}!}{n_{i.}!} \int_{\theta_i \in C} \frac{\prod_{j=1}^{K} \theta_{ij}^{\mu_{(h)}\tau_{(h)}-1}}{\prod_{j=1}^{K} \theta_{ij}^{n_{ij}+\mu_{(h)}\tau_{(h)}-1}} \frac{\prod_{j=1}^{K} \theta_{ij}^{n_{ij}+\mu_{(h)}\tau_{(h)}-1}}{\int_{\theta_i \in C} \prod_{j=1}^{K} \theta_{ij}^{n_{ij}+\mu_{(h)}\tau_{(h)}-1}} d\,\boldsymbol{\theta}_i \right]^{-1},
\end{aligned}$$

where

$$\frac{\prod_{j=1}^{K} \theta_{ij}^{n_{ij}+\mu_{(h)}\tau_{(h)}-1}}{\int_{\theta_i \in C} \prod_{j=1}^{K} \theta_{ij}^{n_{ij}+\mu_{(h)}\tau_{(h)}-1} d\boldsymbol{\theta}_i}$$

is the density function of $\boldsymbol{\theta}_i$, and $\boldsymbol{\theta}_i \in C$.

We notice $\boldsymbol{\mu}^{(h)}$ and $\tau^{(h)}$ are the posterior samples from Section 7.2. For each pair of $\boldsymbol{\mu}^{(h)}$ and $\tau^{(h)}$, we can draw $\boldsymbol{\theta}_i$ from Dirichlet$\left(\mathbf{n}_i + \boldsymbol{\mu}^{(h)}\tau^{(h)}\right)$,

$$\widehat{\mathrm{CPO}}_{i(M_2)} = \left[\frac{1}{M}\sum_{h=1}^{M} \frac{\prod_{j=1}^{K} n_{ij}!}{n_{i.}!}\left(\frac{1}{M'}\sum_{h'=1}^{M'}\prod_{j=1}^{K} \theta_{ij}^{(h')^{-n_{ij}}}\right)\right]^{-1},$$

where $\boldsymbol{\theta}_i^{(h')} \sim$ Dirichlet$\left(\mathbf{n}_i + \boldsymbol{\mu}^{(h)}\tau^{(h)}\right)$ with order restriction. Then we get the LPML as $\widehat{\mathrm{LPML}} = \sum_{i=1}^{I} \log \left(\widehat{\mathrm{CPO}}_i\right)$.

However, it is not easy to compute CPO_i or $\widehat{\mathrm{CPO}}_i$ of M_4 directly. We present how to use the known CPOs, such as $\mathrm{CPO}_{i(M_2)}$ and $\mathrm{CPO}_{i(M_3)}$, to compute $\mathrm{CPO}_{i(M_4)}$,

$$\begin{aligned}
\mathrm{CPO}_{i(M_4)} &= f\left(n_i \,|\, n_{(i)}\right) = \left(\frac{f\left(n_{(i)}\right)}{f(n)}\right)^{-1} \\
&= \left[\frac{\sum_{\ell=1}^{K} P(L=\ell) \iint f\left(n_{(i)} \,|\, \boldsymbol{\mu}, \tau, L=\ell\right) f\left(\boldsymbol{\mu}, \tau \,|\, L=\ell\right) d\boldsymbol{\mu} d\tau}{f(n)}\right]^{-1} \\
&= \left[\sum_{\ell=1}^{K} P(L=\ell) \iint \frac{f\left(n_{(i)} \,|\, \boldsymbol{\mu}, \tau, L=\ell\right) f\left(\boldsymbol{\mu}, \tau \,|\, L=\ell\right)}{f(n)} d\boldsymbol{\mu} d\tau\right]^{-1} \\
&= \left[\sum_{\ell=1}^{K} P(L=\ell) \iint \frac{f\left(n_i \,|\, \boldsymbol{\mu}, \tau, L=\ell\right) f\left(n_{(i)} \,|\, \boldsymbol{\mu}, \tau, L=\ell\right) f\left(\boldsymbol{\mu}, \tau \,|\, L=\ell\right)}{f\left(n_i \,|\, \boldsymbol{\mu}, \tau, L=\ell\right) f(n)} d\boldsymbol{\mu} d\tau\right]^{-1} \\
&= \left[\sum_{\ell=1}^{K} P(L=\ell) \iint \frac{f\left(n \,|\, \boldsymbol{\mu}, \tau, L=\ell\right) f\left(\boldsymbol{\mu}, \tau \,|\, L=\ell\right)}{f\left(n_i \,|\, \boldsymbol{\mu}, \tau, L=\ell\right) f(n)} d\boldsymbol{\mu} d\tau\right]^{-1} \\
&= \left[\sum_{\ell=1}^{K} P(L=\ell) \iint \frac{f\left(n \,|\, L=\ell\right)}{f\left(n_i \,|\, \boldsymbol{\mu}, \tau, L=\ell\right) f(n)} \frac{f\left(n \,|\, \boldsymbol{\mu}, \tau, L=\ell\right) f\left(\boldsymbol{\mu}, \tau \,|\, L=\ell\right)}{f\left(n \,|\, L=\ell\right)} d\boldsymbol{\mu} d\tau\right]^{-1} \\
&= \left[\sum_{\ell=1}^{K} P(L=\ell) \frac{f\left(n \,|\, L=\ell\right)}{f(n)} \iint \frac{1}{f\left(n_i \,|\, \boldsymbol{\mu}, \tau, L=\ell\right)} \frac{f\left(n \,|\, \boldsymbol{\mu}, \tau, L=\ell\right) f\left(\boldsymbol{\mu}, \tau \,|\, L=\ell\right)}{f\left(n \,|\, L=\ell\right)} d\boldsymbol{\mu} d\tau\right]^{-1} \\
&= \left[\sum_{\ell=1}^{K} \frac{P(L=\ell) f\left(n \,|\, L=\ell\right)}{\sum_{\ell=1}^{K} P(L=\ell) f\left(n \,|\, \ell\right)} \iint \frac{1}{f\left(n_i \,|\, \boldsymbol{\mu}, \tau, L=\ell\right)} f\left(\boldsymbol{\mu}, \tau \,|\, n, L=\ell\right) d\boldsymbol{\mu} d\tau\right]^{-1} \\
&= \left[\sum_{\ell=1}^{K} P\left(L=\ell \,|\, n\right) \iint \frac{1}{f\left(n_i \,|\, \boldsymbol{\mu}, \tau, L=\ell\right)} f\left(\boldsymbol{\mu}, \tau \,|\, n, L=\ell\right) d\boldsymbol{\mu} d\tau\right]^{-1},
\end{aligned}$$

then $\widehat{\mathrm{CPO}}_{i\,(M_4)} \approx \left[\sum_{\ell=1}^{K} P\left(\widehat{L=\ell} \,|\, n\right) \frac{1}{\widehat{\mathrm{CPO}}_{i\,(L_{\mathrm{pos}}=\ell)}}\right]^{-1}$, where $\widehat{\mathrm{CPO}}_{i\,(L_{\mathrm{pos}}=\ell)}$ are known, such as $\widehat{\mathrm{CPO}}_{i\,(M_2)}$ and $\widehat{\mathrm{CPO}}_{i\,(M_3)}$. Without extra computation, taking advantage of known CPOs from M_2 and M_3, we can easily acquire the CPO of M_4.

A.4 Posterior summary of θ

Table A.1
Part I: Counties 1-11

County ID	Model	Underweight			Normal			Overweight			Obese I			Obese II		
		PM	PSD	CV	PM	PSD	CV	PM	PSD	CV	PM	PSD	CV	PM	PSD	CV
1	M_1	0.026	0.013	0.501	0.399	0.040	0.101	0.394	0.040	0.102	0.143	0.029	0.206	0.039	0.016	0.408
	M_2	0.021	0.009	0.425	0.421	0.023	0.056	0.376	0.021	0.056	0.148	0.023	0.153	0.033	0.010	0.316
	M_3	0.021	0.009	0.431	0.376	0.019	0.051	0.418	0.023	0.055	0.152	0.023	0.153	0.033	0.011	0.323
	M_4	0.021	0.009	0.431	0.393	0.030	0.076	0.404	0.030	0.075	0.150	0.023	0.156	0.033	0.010	0.315
2	M_1	0.014	0.010	0.704	0.390	0.040	0.102	0.417	0.041	0.098	0.160	0.030	0.189	0.019	0.011	0.580
	M_2	0.015	0.007	0.490	0.422	0.024	0.056	0.381	0.019	0.049	0.159	0.024	0.152	0.023	0.009	0.386
	M_3	0.015	0.007	0.494	0.375	0.020	0.055	0.426	0.025	0.059	0.161	0.023	0.143	0.023	0.010	0.405
	M_4	0.015	0.007	0.476	0.391	0.031	0.079	0.409	0.031	0.077	0.161	0.024	0.147	0.024	0.010	0.405
3	M_1	0.028	0.014	0.489	0.282	0.039	0.137	0.495	0.042	0.085	0.149	0.029	0.192	0.047	0.017	0.368
	M_2	0.024	0.011	0.459	0.393	0.021	0.054	0.378	0.018	0.047	0.166	0.028	0.167	0.040	0.015	0.368
	M_3	0.021	0.009	0.440	0.334	0.035	0.106	0.458	0.036	0.079	0.151	0.022	0.146	0.037	0.012	0.320
	M_4	0.022	0.010	0.452	0.354	0.042	0.118	0.429	0.050	0.117	0.156	0.026	0.163	0.038	0.013	0.342
4	M_1	0.007	0.004	0.543	0.356	0.022	0.062	0.421	0.022	0.053	0.183	0.018	0.096	0.034	0.009	0.252
	M_2	0.009	0.004	0.461	0.394	0.014	0.035	0.381	0.011	0.029	0.182	0.020	0.112	0.034	0.008	0.224
	M_3	0.009	0.004	0.451	0.363	0.018	0.050	0.422	0.019	0.046	0.174	0.017	0.098	0.032	0.007	0.220
	M_4	0.009	0.004	0.456	0.374	0.023	0.061	0.407	0.026	0.063	0.177	0.018	0.104	0.032	0.007	0.221
5	M_1	0.016	0.011	0.708	0.370	0.042	0.112	0.400	0.042	0.104	0.180	0.033	0.181	0.035	0.016	0.453
	M_2	0.015	0.008	0.515	0.413	0.024	0.057	0.372	0.021	0.057	0.168	0.027	0.158	0.032	0.012	0.360
	M_3	0.015	0.007	0.490	0.366	0.023	0.063	0.419	0.027	0.063	0.169	0.026	0.152	0.032	0.011	0.341
	M_4	0.015	0.008	0.493	0.382	0.032	0.084	0.402	0.033	0.083	0.169	0.026	0.154	0.032	0.011	0.356
6	M_1	0.009	0.009	0.943	0.380	0.045	0.118	0.402	0.044	0.108	0.147	0.032	0.217	0.063	0.021	0.339
	M_2	0.012	0.007	0.586	0.417	0.025	0.059	0.375	0.020	0.054	0.151	0.024	0.160	0.046	0.017	0.362
	M_3	0.012	0.007	0.569	0.371	0.023	0.061	0.423	0.026	0.061	0.151	0.023	0.150	0.043	0.015	0.355
	M_4	0.012	0.007	0.590	0.387	0.032	0.083	0.406	0.034	0.083	0.151	0.024	0.158	0.044	0.016	0.370
7	M_1	0.009	0.009	0.943	0.376	0.044	0.117	0.400	0.045	0.113	0.183	0.035	0.191	0.032	0.016	0.502
	M_2	0.012	0.007	0.575	0.416	0.025	0.059	0.374	0.022	0.058	0.169	0.028	0.163	0.030	0.012	0.389
	M_3	0.013	0.007	0.578	0.367	0.023	0.062	0.422	0.027	0.065	0.169	0.025	0.150	0.030	0.011	0.359
	M_4	0.012	0.007	0.590	0.384	0.033	0.087	0.405	0.034	0.084	0.169	0.027	0.156	0.030	0.011	0.372
8	M_1	0.019	0.014	0.726	0.387	0.048	0.123	0.443	0.050	0.112	0.126	0.033	0.265	0.025	0.015	0.597
	M_2	0.017	0.009	0.520	0.426	0.025	0.058	0.386	0.020	0.051	0.143	0.024	0.170	0.027	0.011	0.406
	M_3	0.016	0.008	0.488	0.376	0.023	0.061	0.437	0.029	0.066	0.144	0.023	0.160	0.027	0.010	0.387
	M_4	0.017	0.009	0.520	0.394	0.035	0.088	0.418	0.035	0.083	0.144	0.023	0.162	0.027	0.011	0.401
9	M_1	0.016	0.011	0.686	0.391	0.045	0.116	0.398	0.044	0.110	0.174	0.035	0.203	0.021	0.012	0.584
	M_2	0.015	0.008	0.504	0.421	0.027	0.064	0.373	0.021	0.058	0.165	0.025	0.152	0.026	0.010	0.389
	M_3	0.016	0.008	0.492	0.372	0.021	0.056	0.420	0.025	0.059	0.167	0.025	0.149	0.025	0.010	0.389
	M_4	0.015	0.008	0.496	0.390	0.033	0.084	0.403	0.033	0.081	0.166	0.025	0.148	0.026	0.010	0.383
10	M_1	0.008	0.007	0.940	0.396	0.041	0.103	0.403	0.042	0.104	0.180	0.033	0.184	0.013	0.010	0.760
	M_2	0.011	0.007	0.574	0.423	0.024	0.057	0.377	0.022	0.058	0.167	0.025	0.151	0.021	0.010	0.453
	M_3	0.012	0.007	0.573	0.376	0.021	0.055	0.422	0.024	0.057	0.169	0.025	0.146	0.021	0.009	0.438
	M_4	0.012	0.007	0.579	0.393	0.033	0.083	0.406	0.032	0.079	0.168	0.025	0.146	0.021	0.009	0.447
11	M_1	0.026	0.013	0.515	0.365	0.037	0.102	0.385	0.038	0.098	0.181	0.030	0.167	0.044	0.016	0.366
	M_2	0.021	0.009	0.420	0.407	0.024	0.058	0.367	0.021	0.057	0.169	0.025	0.148	0.036	0.012	0.323
	M_3	0.021	0.009	0.435	0.363	0.022	0.062	0.411	0.026	0.064	0.169	0.024	0.144	0.037	0.012	0.326
	M_4	0.021	0.009	0.440	0.379	0.031	0.081	0.395	0.031	0.078	0.169	0.024	0.140	0.036	0.012	0.322

Note: Posterior Mean (PM), Posterior Standard Deviation (PSD), Coefficient of Variation (CV).

Table A.2
Part II: Counties 12-23

County ID	Model	Underweight			Normal			Overweight			Obese I			Obese II		
		PM	PSD	CV	PM	PSD	CV	PM	PSD	CV	PM	PSD	CV	PM	PSD	CV
12	M_1	0.008	0.007	0.937	0.415	0.041	0.099	0.439	0.042	0.095	0.113	0.027	0.235	0.026	0.013	0.507
	M_2	0.012	0.007	0.581	0.434	0.024	0.055	0.392	0.020	0.050	0.135	0.023	0.171	0.028	0.010	0.360
	M_3	0.012	0.007	0.557	0.386	0.022	0.056	0.438	0.026	0.059	0.137	0.024	0.173	0.027	0.010	0.355
	M_4	0.012	0.007	0.583	0.403	0.033	0.082	0.422	0.033	0.078	0.135	0.024	0.176	0.028	0.010	0.357
13	M_1	0.012	0.007	0.563	0.432	0.030	0.070	0.378	0.029	0.076	0.142	0.021	0.146	0.036	0.012	0.323
	M_2	0.013	0.006	0.426	0.434	0.023	0.053	0.375	0.020	0.053	0.146	0.018	0.123	0.033	0.009	0.272
	M_3	0.013	0.006	0.423	0.388	0.014	0.037	0.413	0.017	0.042	0.152	0.019	0.122	0.034	0.009	0.277
	M_4	0.013	0.006	0.426	0.405	0.028	0.069	0.399	0.025	0.063	0.150	0.019	0.124	0.033	0.009	0.273
14	M_1	0.024	0.013	0.545	0.425	0.045	0.106	0.399	0.044	0.110	0.131	0.030	0.228	0.022	0.012	0.567
	M_2	0.019	0.009	0.465	0.434	0.027	0.062	0.378	0.023	0.059	0.144	0.023	0.162	0.025	0.010	0.380
	M_3	0.019	0.009	0.463	0.383	0.021	0.055	0.426	0.024	0.057	0.147	0.024	0.162	0.026	0.010	0.389
	M_4	0.019	0.009	0.465	0.400	0.033	0.082	0.409	0.032	0.078	0.146	0.024	0.162	0.025	0.010	0.378
15	M_1	0.022	0.012	0.532	0.357	0.041	0.114	0.444	0.041	0.093	0.131	0.028	0.214	0.047	0.018	0.384
	M_2	0.018	0.008	0.438	0.412	0.021	0.050	0.384	0.017	0.045	0.148	0.025	0.166	0.039	0.013	0.334
	M_3	0.018	0.008	0.462	0.368	0.025	0.068	0.433	0.028	0.064	0.145	0.023	0.155	0.037	0.012	0.325
	M_4	0.018	0.008	0.448	0.383	0.032	0.083	0.416	0.035	0.083	0.146	0.024	0.167	0.037	0.012	0.327
16	M_1	0.013	0.009	0.695	0.372	0.037	0.100	0.439	0.041	0.092	0.158	0.029	0.183	0.018	0.010	0.584
	M_2	0.015	0.007	0.482	0.416	0.020	0.048	0.386	0.017	0.044	0.160	0.024	0.150	0.023	0.009	0.406
	M_3	0.014	0.007	0.480	0.371	0.023	0.062	0.436	0.028	0.063	0.157	0.021	0.135	0.023	0.009	0.383
	M_4	0.014	0.007	0.481	0.386	0.031	0.080	0.418	0.035	0.083	0.158	0.023	0.147	0.023	0.009	0.381
17	M_1	0.039	0.016	0.405	0.351	0.039	0.111	0.426	0.041	0.095	0.161	0.030	0.187	0.024	0.012	0.507
	M_2	0.028	0.012	0.418	0.406	0.021	0.051	0.378	0.017	0.045	0.161	0.025	0.153	0.027	0.010	0.362
	M_3	0.026	0.011	0.420	0.362	0.024	0.066	0.428	0.028	0.064	0.157	0.021	0.132	0.027	0.009	0.351
	M_4	0.027	0.012	0.425	0.377	0.030	0.080	0.410	0.034	0.083	0.159	0.023	0.142	0.027	0.010	0.365
18	M_1	0.009	0.009	0.964	0.420	0.045	0.108	0.376	0.043	0.114	0.164	0.036	0.220	0.032	0.017	0.519
	M_2	0.012	0.007	0.581	0.430	0.028	0.065	0.370	0.024	0.066	0.158	0.026	0.163	0.030	0.011	0.373
	M_3	0.013	0.007	0.552	0.378	0.019	0.051	0.417	0.024	0.056	0.162	0.025	0.153	0.031	0.011	0.362
	M_4	0.013	0.007	0.568	0.396	0.034	0.086	0.400	0.033	0.082	0.161	0.025	0.159	0.031	0.011	0.366
19	M_1	0.019	0.013	0.693	0.416	0.048	0.116	0.384	0.047	0.123	0.164	0.035	0.214	0.016	0.012	0.767
	M_2	0.016	0.008	0.507	0.431	0.030	0.070	0.372	0.025	0.066	0.157	0.026	0.162	0.023	0.010	0.430
	M_3	0.017	0.009	0.532	0.378	0.020	0.053	0.420	0.025	0.059	0.162	0.025	0.158	0.024	0.010	0.407
	M_4	0.017	0.009	0.533	0.397	0.036	0.091	0.402	0.034	0.085	0.161	0.027	0.166	0.024	0.010	0.422
20	M_1	0.009	0.009	0.935	0.335	0.044	0.132	0.494	0.047	0.095	0.139	0.031	0.225	0.023	0.013	0.564
	M_2	0.013	0.008	0.610	0.413	0.020	0.048	0.390	0.017	0.043	0.157	0.027	0.171	0.027	0.011	0.406
	M_3	0.012	0.007	0.551	0.359	0.029	0.082	0.454	0.035	0.077	0.149	0.023	0.156	0.026	0.010	0.380
	M_4	0.012	0.007	0.599	0.378	0.037	0.098	0.432	0.043	0.100	0.152	0.025	0.166	0.026	0.010	0.396
21	M_1	0.048	0.021	0.431	0.431	0.050	0.116	0.353	0.051	0.145	0.123	0.033	0.269	0.046	0.021	0.453
	M_2	0.029	0.012	0.432	0.436	0.032	0.074	0.363	0.029	0.079	0.138	0.025	0.179	0.035	0.013	0.363
	M_3	0.029	0.014	0.485	0.377	0.020	0.052	0.412	0.024	0.058	0.146	0.025	0.174	0.036	0.013	0.364
	M_4	0.029	0.014	0.459	0.398	0.038	0.096	0.394	0.035	0.090	0.143	0.026	0.180	0.036	0.013	0.372
22	M_1	0.016	0.010	0.660	0.431	0.044	0.102	0.391	0.043	0.109	0.134	0.030	0.226	0.029	0.015	0.512
	M_2	0.015	0.008	0.500	0.434	0.027	0.062	0.378	0.023	0.060	0.145	0.024	0.163	0.028	0.010	0.369
	M_3	0.015	0.008	0.500	0.384	0.019	0.050	0.423	0.023	0.055	0.149	0.023	0.151	0.029	0.011	0.362
	M_4	0.015	0.008	0.508	0.402	0.034	0.083	0.407	0.032	0.078	0.147	0.024	0.160	0.029	0.011	0.376
23	M_1	0.011	0.011	0.979	0.379	0.048	0.126	0.426	0.048	0.112	0.149	0.034	0.230	0.035	0.018	0.516
	M_2	0.013	0.007	0.560	0.422	0.025	0.060	0.379	0.021	0.055	0.155	0.026	0.171	0.031	0.011	0.352
	M_3	0.013	0.007	0.568	0.371	0.024	0.064	0.431	0.029	0.068	0.154	0.025	0.162	0.032	0.012	0.378
	M_4	0.013	0.007	0.570	0.388	0.035	0.089	0.413	0.037	0.089	0.155	0.026	0.171	0.032	0.012	0.365

Note: Posterior Mean (PM), Posterior Standard Deviation (PSD), Coefficient of Variation (CV).

Table A.3
Part III: Counties 24-35

County ID	Model	Underweight			Normal			Overweight			Obese I			Obese II		
		PM	PSD	CV	PM	PSD	CV	PM	PSD	CV	PM	PSD	CV	PM	PSD	CV
24	M_1	0.008	0.008	1.005	0.375	0.044	0.116	0.397	0.043	0.107	0.182	0.034	0.189	0.038	0.017	0.445
	M_2	0.012	0.007	0.596	0.414	0.024	0.058	0.373	0.021	0.055	0.167	0.027	0.160	0.033	0.011	0.339
	M_3	0.012	0.007	0.551	0.368	0.023	0.062	0.418	0.026	0.061	0.169	0.025	0.145	0.033	0.011	0.339
	M_4	0.012	0.007	0.581	0.385	0.033	0.085	0.403	0.032	0.079	0.168	0.026	0.153	0.032	0.011	0.343
25	M_1	0.018	0.012	0.676	0.449	0.047	0.103	0.402	0.045	0.112	0.117	0.029	0.248	0.015	0.011	0.751
	M_2	0.016	0.008	0.483	0.444	0.030	0.068	0.383	0.023	0.060	0.135	0.025	0.185	0.022	0.010	0.435
	M_3	0.016	0.008	0.512	0.390	0.020	0.050	0.428	0.024	0.055	0.143	0.025	0.177	0.023	0.010	0.422
	M_4	0.016	0.008	0.510	0.411	0.036	0.087	0.412	0.033	0.080	0.139	0.026	0.188	0.023	0.009	0.421
26	M_1	0.027	0.016	0.595	0.373	0.045	0.120	0.432	0.046	0.107	0.136	0.032	0.232	0.032	0.016	0.514
	M_2	0.021	0.010	0.483	0.417	0.023	0.056	0.383	0.019	0.050	0.148	0.026	0.173	0.031	0.012	0.378
	M_3	0.020	0.009	0.477	0.370	0.025	0.066	0.433	0.029	0.066	0.148	0.024	0.161	0.029	0.010	0.357
	M_4	0.020	0.009	0.463	0.387	0.034	0.087	0.415	0.035	0.084	0.148	0.025	0.168	0.030	0.011	0.365
27	M_1	0.030	0.018	0.582	0.302	0.045	0.148	0.473	0.049	0.103	0.170	0.037	0.219	0.026	0.016	0.600
	M_2	0.022	0.011	0.492	0.401	0.023	0.056	0.378	0.019	0.050	0.171	0.030	0.176	0.028	0.011	0.377
	M_3	0.020	0.009	0.463	0.346	0.034	0.099	0.446	0.037	0.082	0.160	0.024	0.150	0.027	0.011	0.386
	M_4	0.021	0.010	0.479	0.366	0.041	0.112	0.423	0.046	0.109	0.163	0.027	0.163	0.028	0.011	0.391
28	M_1	0.019	0.013	0.687	0.410	0.047	0.115	0.389	0.048	0.122	0.156	0.035	0.221	0.025	0.015	0.594
	M_2	0.017	0.008	0.494	0.429	0.028	0.066	0.374	0.025	0.066	0.154	0.026	0.168	0.027	0.010	0.389
	M_3	0.017	0.008	0.504	0.377	0.022	0.058	0.421	0.025	0.059	0.159	0.027	0.167	0.027	0.010	0.373
	M_4	0.017	0.009	0.508	0.395	0.034	0.087	0.404	0.035	0.086	0.157	0.026	0.168	0.027	0.011	0.394
29	M_1	0.009	0.008	0.980	0.391	0.042	0.107	0.429	0.041	0.096	0.150	0.032	0.211	0.022	0.013	0.575
	M_2	0.012	0.007	0.621	0.424	0.023	0.055	0.384	0.020	0.051	0.155	0.024	0.156	0.025	0.010	0.394
	M_3	0.012	0.007	0.566	0.376	0.023	0.060	0.433	0.027	0.062	0.154	0.023	0.147	0.025	0.009	0.370
	M_4	0.012	0.007	0.591	0.393	0.033	0.083	0.416	0.033	0.081	0.155	0.023	0.149	0.025	0.009	0.372
30	M_1	0.015	0.010	0.702	0.338	0.041	0.121	0.420	0.044	0.104	0.207	0.034	0.166	0.020	0.012	0.590
	M_2	0.016	0.007	0.471	0.401	0.022	0.055	0.373	0.019	0.052	0.186	0.032	0.171	0.025	0.010	0.380
	M_3	0.015	0.007	0.466	0.355	0.027	0.075	0.427	0.028	0.066	0.179	0.028	0.155	0.024	0.009	0.386
	M_4	0.015	0.007	0.468	0.371	0.033	0.090	0.407	0.037	0.090	0.183	0.030	0.165	0.025	0.009	0.386
31	M_1	0.023	0.013	0.578	0.399	0.043	0.107	0.391	0.043	0.110	0.158	0.031	0.199	0.030	0.015	0.491
	M_2	0.019	0.009	0.462	0.423	0.026	0.062	0.373	0.022	0.060	0.156	0.025	0.161	0.029	0.011	0.374
	M_3	0.019	0.009	0.478	0.373	0.022	0.058	0.420	0.025	0.060	0.160	0.025	0.155	0.028	0.010	0.351
	M_4	0.019	0.009	0.472	0.391	0.033	0.083	0.403	0.033	0.082	0.159	0.025	0.158	0.029	0.010	0.355
32	M_1	0.007	0.007	0.941	0.319	0.037	0.116	0.450	0.039	0.086	0.200	0.032	0.159	0.024	0.012	0.511
	M_2	0.012	0.007	0.569	0.397	0.020	0.051	0.378	0.016	0.042	0.186	0.031	0.164	0.027	0.010	0.370
	M_3	0.011	0.006	0.576	0.348	0.029	0.084	0.439	0.030	0.068	0.177	0.026	0.144	0.026	0.009	0.345
	M_4	0.011	0.006	0.579	0.365	0.036	0.097	0.417	0.039	0.094	0.181	0.029	0.159	0.026	0.009	0.352
33	M_1	0.011	0.007	0.662	0.367	0.037	0.101	0.419	0.035	0.084	0.177	0.029	0.164	0.026	0.012	0.458
	M_2	0.014	0.007	0.510	0.411	0.020	0.049	0.381	0.017	0.044	0.168	0.024	0.140	0.027	0.009	0.331
	M_3	0.013	0.006	0.502	0.370	0.021	0.058	0.424	0.024	0.056	0.167	0.022	0.133	0.027	0.009	0.346
	M_4	0.013	0.007	0.519	0.384	0.029	0.076	0.408	0.031	0.076	0.169	0.023	0.135	0.027	0.009	0.352
34	M_1	0.015	0.010	0.695	0.373	0.041	0.110	0.452	0.042	0.092	0.134	0.030	0.222	0.026	0.013	0.503
	M_2	0.015	0.008	0.496	0.420	0.021	0.051	0.389	0.017	0.044	0.148	0.023	0.158	0.028	0.011	0.390
	M_3	0.015	0.007	0.485	0.372	0.024	0.065	0.443	0.029	0.065	0.144	0.022	0.153	0.027	0.010	0.363
	M_4	0.015	0.007	0.495	0.388	0.033	0.086	0.424	0.036	0.085	0.145	0.023	0.157	0.028	0.011	0.381
35	M_1	0.014	0.010	0.705	0.419	0.040	0.095	0.435	0.040	0.092	0.121	0.028	0.228	0.012	0.010	0.790
	M_2	0.015	0.007	0.488	0.436	0.024	0.055	0.392	0.020	0.050	0.138	0.022	0.162	0.020	0.009	0.447
	M_3	0.014	0.007	0.474	0.388	0.021	0.055	0.437	0.026	0.059	0.140	0.023	0.166	0.020	0.009	0.433
	M_4	0.015	0.007	0.486	0.406	0.032	0.080	0.421	0.033	0.077	0.139	0.023	0.167	0.020	0.009	0.439

Note: Posterior Mean (PM), Posterior Standard Deviation (PSD), Coefficient of Variation (CV).

References

Chen, X., and Nandram, B. (2019). Order-restricted Bayesian estimation of multinomial cell counts for small areas. In *JSM 2019*, Section on Bayesian Statistical Science, Alexandria, VA: American Statistical Association, 1024-1030.

Chen, X., and Nandram, B. (2021). Bayesian order-restricted inference of multinomial counts from small areas. *Recent Advances in Applied Statistics*. David Hangal, Raosahed Laptate, Hukum Chandra and Girish Chandra, Springer Nature (in press).

Chen, M.-H., and Shao, Q.-M. (1997). On Monte Carlo methods for estimating ratios of normalizing constants. *The Annals of Statistics*, 25(4), 1563-1594.

Gelfand, A.E., Dey, D.K. and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. Technical report, Stanford University, Department of Statistics.

Gelfand, A.E., Smith, A.F.M. and Lee, T.-M. (1992). Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *Journal of the American Statistical Association*, 87(418), 523-532.

Geweke, J. (1991). *Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments*, Federal Reserve Bank of Minneapolis, Research Department Minneapolis, MN, 196.

Ghosh, M., and Rao, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9(1), 55-76.

Heck, D.W., and Davis-Stober, C.P. (2019). Multinomial models with linear inequality constraints: Overview and improvements of computational methods for Bayesian inference. *Journal of Mathematical Psychology*, 91, 70-87.

Liu, J., and Sabatti, C. (2000). Generalised Gibbs sampler and multigrid Monte Carlo for Bayesian computation. *Biometrika*, 87(2), 353-369.

Mair, P., Hornik, K. and de Leeuw, J. (2009). Isotone optimization in r: pool-adjacent-violators algorithm (pava) and active set methods. *Journal of Statistical Software*, 32(5), 1-24.

Maples, J. (2019). Small area estimates for child population and poverty in school districts using dirichlet-multinomial models. In *JSM 2019*. Alexandria, VA: American Statistical Association.

Nadarajah, S., and Kotz, S. (2006). R programs for truncated distributions. *Journal of Statistical Software*, 16(1), 1-8.

Nandram, B. (1997). Bayesian inference for the best ordinal multinomial population in a taste test. In *Case Studies in Bayesian Statistics*, (Eds., C. Gatsonis, J.S. Hodges, R.E. Kass, R. McCulloch, P. Rossi and N.D. Singpurwalla), New York: Springer, 399-418.

Nandram, B. (1998). A Bayesian analysis of the three-stage hierarchical multinomial model. *Journal of Statistical Computation and Simulation*, 61(1-2), 97-126.

Nandram, B., and Sedransk, J. (1995). Bayesian inference for the mean of a stratified population when there are order restrictions. In *Case Studies in Bayesian Statistics, Volume II*, Springer, 309-322.

Nandram, B., Choi, J.W. and Liu, Y. (2021). Integration of nonprobability and probability samples via survey weights. *International Journal of Statistics and Probability*, 10(6), 5-21.

Nandram, B., Erciulescu, A.L. and Cruze, N.B. (2019). Bayesian benchmarking of the Fay-Herriot model using random deletion. *Survey Methodology*, 45, 2, 365-390. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2019002/article/00004-eng.pdf.

Nandram, B., Kim, D. and Zhou, J. (2019). A pooled Bayes test of independence for sparse contingency tables from small areas. *Journal of Statistical Computation and Simulation*, 89(5), 899-926.

Nandram, B., Sedransk, J. and Smith, S.J. (1997). Order-restricted Bayesian estimation of the age composition of a population of Atlantic cod. *Journal of the American Statistical Association*, 92(437), 33-40.

Neuenschwander, B., Wandel, S., Roychoudhury, S. and Bailey, S. (2016). Robust exchangeability designs for early phase clinical trials with multiple strata. *Pharmaceutical Statistics*, 15(2), 123-134.

Rao, J., and Molina, I. (2015). *Small Area Estimation*. Wiley Series in Survey Methodology, Wiley.

Sedransk, J., Monahan, J. and Chiu, H.Y. (1985). Bayesian estimation of finite population parameters in categorical data models incorporating order restrictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 47(3), 519-527.

Trevisani, M., and Torelli, N. (2007). Hierarchical Bayesian models for small area estimation with count data. *Open Journal of Statistics*, 07.

Wang, X., Berg, E., Zhu, Z., Sun, D. and Demuth, G. (2018). Small area estimation of proportions with constraint for national resources inventory survey. *Journal of Agricultural, Biological and Environmental Statistics*, 23(4), 509-528.

Wu, J., Meyer, M.C. and Opsomer, J.D. (2016). Survey estimation of domain means that respect natural orderings. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, 44(4), 431-444.

Yang, L. (2021). Bayesian predictive inference with survey weights. Master's thesis, Department of Mathematical Science, Worcester Polytechnic Institute.

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

Survey Methodology, June 2022
Vol. 48, No. 1, pp. 177-190
Statistics Canada, Catalogue No. 12-001-X

A generalization of inverse probability weighting

Alain Théberge[1]

Abstract

In finite population estimation, the inverse probability or Horvitz-Thompson estimator is a basic tool. Even when auxiliary information is available to model the variable of interest, it is still used to estimate the model error. Here, the inverse probability estimator is generalized by introducing a positive definite matrix. The usual inverse probability estimator is a special case of the generalized estimator, where the positive definite matrix is the identity matrix. Since calibration estimation seeks weights that are close to the inverse probability weights, it too can be generalized by seeking weights that are close to those of the generalized inverse probability estimator. Calibration is known to be optimal, in the sense that it asymptotically attains the Godambe-Joshi lower bound. That lower bound has been derived under a model where no correlation is present. This too, can be generalized to allow for correlation. With the correct choice of the positive definite matrix that generalizes the calibration estimators, this generalized lower bound can be asymptotically attained. There is often no closed-form formula for the generalized estimators. However, simple explicit examples are given here to illustrate how the generalized estimators take advantage of the correlation. This simplicity is achieved here, by assuming a correlation of one between some population units. Those simple estimators can still be useful, even if the correlation is smaller than one. Simulation results are used to compare the generalized estimators to the ordinary estimators.

Key Words: Calibration estimator; Godambe-Joshi lower bound; Horvitz-Thompson estimator; Moore-Penrose inverse; Vaccination rate.

1. Introduction

The usual inverse probability estimator of the total for a population of N units is

$$\hat{\theta}_{\mathrm{IP}} = \sum_{i=1}^{N} \frac{\delta_i y_i}{\pi_i}, \tag{1.1}$$

where y_i is the variable of interest for unit i, δ_i is 1 or 0 depending on whether i is in the sample s or not, and $\pi_i > 0$ is the probability that i is in s. Note that the expectation of δ_i is π_i, this makes $\hat{\theta}_{\mathrm{IP}}$ unbiased for $\theta = \sum_{i=1}^{N} y_i$. It is also known as the Horvitz-Thompson estimator, presented in Horvitz and Thompson (1952). In this paper, estimators that can draw some strength from units not in s will be presented.

Here is an example of such an estimator for a population of N units that is partitioned into $N_p = N/2$ pairs $\{2i-1, 2i\}$ $(i = 1, 2, \ldots, N_p)$,

$$\hat{\theta}_{\mathrm{LIM}} = \sum_{i=1}^{N_p} \frac{2y_{2i-1}\delta_{2i-1} + 2y_{2i}\delta_{2i} - \left(y_{2i-1} + y_{2i}\right)\delta_{2i-1}\delta_{2i} + \left(y_{2i-1} - y_{2i}\right)\delta_{2i-1}\delta_{2i}\,\pi_{\mathrm{diff}\ i}}{\pi_{2i-1} + \pi_{2i} - \pi_{2i-1\ 2i}}, \tag{1.2}$$

where $\pi_{2i-1\ 2i} = E\left(\delta_{2i-1}\delta_{2i}\right)$ is the probability that both units $2i-1$ and $2i$ are in s, and $\pi_{\mathrm{diff}\ i} = \left(\pi_{2i} - \pi_{2i-1}\right) \big/ \pi_{2i-1\ 2i}$. It can be verified that $\hat{\theta}_{\mathrm{LIM}}$ is also unbiased.

1. Alain Théberge, Ottawa, Ontario, Canada, K4C 1E2. E-mail: alain.theberge1@gmail.com.

It should be noted that the denominators in (1.2) correspond to the probability that at least one unit of the pair is in the sample. Thus, this estimator is reminiscent of inverse probability weighting, except it is based on pairs, instead of individual units. The numerators in (1.2) correspond to a value assigned to each pair with at least one sampled unit, and each observed pair is given a weight equal to the inverse of the probability of being observed. From the observation of only one unit of a pair, the estimator (1.2) assigns a value to the pair, and if the units of a pair are strongly correlated, this may be an efficient way to utilize this correlation. The estimator is a special case of a more general one that applies to more general populations, not only those with units grouped in pairs. Because it yields examples that give some insight into the general estimator, and because those examples can be given an explicit form that is simple to interpret and understand, Section 6 and Section 7 will also be about the case where the population, or a domain, is partitioned into pairs. The generalized inverse probability estimator is presented in Section 2; it depends on a parameter $\mathbf{\Sigma}$, a positive definite $N \times N$ matrix. In Section 3, the new estimator is applied to the problem of calibration. The choice of the parameter $\mathbf{\Sigma}$ is discussed in Section 4. In Section 5, it is seen that, with the right choice for $\mathbf{\Sigma}$, the generalized calibration estimator is optimal, in the sense that it asymptotically attains a generalization of the Godambe-Joshi lower bound. Simple examples are given in Section 6, and the results of a simulation are presented in Section 7. Section 8 summarizes the paper.

2. The generalized inverse probability estimator

Estimators in this paper utilize a positive definite matrix $\mathbf{\Sigma} \in \mathbb{R}^{N \times N}$. A matrix formulation of the estimators will therefore be useful. For a vector of interest $\mathbf{y}' = (y_1, y_2, \ldots, y_N)$ and $\mathbf{1}_{N\times 1}$ a vector of ones, the inverse probability estimator of the total $\theta = \sum_{i=1}^{N} y_i = \mathbf{y}'\mathbf{1}_{N\times 1}$ can be written

$$\begin{aligned} \hat{\theta}_{\mathrm{IP}} &= \sum_{i=1}^{N} \frac{\delta_i y_i}{\pi_i} \\ &= \mathbf{y}'\mathbf{\Delta}_s \left(E\left(\mathbf{\Delta}_s\right)\right)^{-1} \mathbf{1}_{N\times 1}, \end{aligned} \tag{2.1}$$

where $\pi_i = E(\delta_i)$ is assumed greater than 0 for $i = 1, 2, \ldots, N$, and $\mathbf{\Delta}_s$ is the $N \times N$ diagonal matrix of the δ_i.

The generalization of the inverse probability estimator relies on the Moore-Penrose inverse of a matrix $\mathbf{M}$, denoted $\mathbf{M}^{\dagger}$. The Moore-Penrose inverse is unique and always exists; it is equal to the ordinary inverse if the latter exists. A precise definition and properties of the Moore-Penrose inverse can be found in Ben-Israel and Greville (2002). In particular, it can be verified that $\mathbf{\Delta}_s^{\dagger} = \mathbf{\Delta}_s$. Since it is also true that $\mathbf{\Delta}_s^2 = \mathbf{\Delta}_s$, if $\mathbf{I} \in \mathbb{R}^{N \times N}$ is the identity matrix, the inverse probability estimator can be written

$$\begin{aligned} \hat{\theta}_{\mathrm{IP}} &= \mathbf{y}'\mathbf{\Delta}_s \left(E\left(\mathbf{\Delta}_s\right)\right)^{-1} \mathbf{1}_{N\times 1} \\ &= \mathbf{y}'\left(\mathbf{\Delta}_s \mathbf{I} \mathbf{\Delta}_s\right)^{\dagger} \left(E\left(\mathbf{\Delta}_s \mathbf{I} \mathbf{\Delta}_s\right)^{\dagger}\right)^{-1} \mathbf{1}_{N\times 1}. \end{aligned} \tag{2.2}$$

If in (2.2), the identity matrix is replaced by any $N \times N$ positive definite matrix $\boldsymbol{\Sigma}$, one obtains the generalized inverse probability estimator or the generalized Horvitz-Thompson estimator,

$$\hat{\theta}_{\text{GIP}}(\boldsymbol{\Sigma}) = \mathbf{y}'(\boldsymbol{\Delta}_s\boldsymbol{\Sigma}\boldsymbol{\Delta}_s)^{\dagger}\left(E(\boldsymbol{\Delta}_s\boldsymbol{\Sigma}\boldsymbol{\Delta}_s)^{\dagger}\right)^{-1}\mathbf{1}_{N\times 1}. \tag{2.3}$$

In the phrase "inverse probability", the matrix $E(\boldsymbol{\Delta}_s\boldsymbol{\Sigma}\boldsymbol{\Delta}_s)^{\dagger}$ is now the "probability" and $\left(E(\boldsymbol{\Delta}_s\boldsymbol{\Sigma}\boldsymbol{\Delta}_s)^{\dagger}\right)^{-1}$ is the new "inverse probability". The ordinary inverse probability estimator is simply a special case of $\hat{\theta}_{\text{GIP}}(\boldsymbol{\Sigma})$, which can be obtained by choosing $\boldsymbol{\Sigma} = \mathbf{I}$. As will be seen in c) below, one now has a family of unbiased estimators, $\hat{\theta}_{\text{GIP}}(\boldsymbol{\Sigma})$, parameterized by $\boldsymbol{\Sigma}$.

2.1 Notes on the generalized inverse probability estimator

a) Although the vector $\mathbf{y}$ appears in the estimator, only the sampled units affect the estimator's value. This is because $(\boldsymbol{\Delta}_s\boldsymbol{\Sigma}\boldsymbol{\Delta}_s)^{\dagger} = \boldsymbol{\Delta}_s(\boldsymbol{\Delta}_s\boldsymbol{\Sigma}\boldsymbol{\Delta}_s)^{\dagger}$, thus (2.3) could have been written

$$\hat{\theta}_{\text{GIP}}(\boldsymbol{\Sigma}) = (\boldsymbol{\Delta}_s\mathbf{y})'(\boldsymbol{\Delta}_s\boldsymbol{\Sigma}\boldsymbol{\Delta}_s)^{\dagger}\left(E(\boldsymbol{\Delta}_s\boldsymbol{\Sigma}\boldsymbol{\Delta}_s)^{\dagger}\right)^{-1}\mathbf{1}_{N\times 1}. \tag{2.4}$$

The proof of this and of many other results stated here may be found in Théberge (2017). The $N \times 1$ vector $\mathbf{w}_{s\,\text{GIP}}(\boldsymbol{\Sigma}) = (\boldsymbol{\Delta}_s\boldsymbol{\Sigma}\boldsymbol{\Delta}_s)^{\dagger}\left(E(\boldsymbol{\Delta}_s\boldsymbol{\Sigma}\boldsymbol{\Delta}_s)^{\dagger}\right)^{-1}\mathbf{1}_{N\times 1}$ gives the weights of $\hat{\theta}_{\text{GIP}}(\boldsymbol{\Sigma})$, and all the units not in sample have a weight of zero.

b) The matrix $E(\boldsymbol{\Delta}_s\boldsymbol{\Sigma}\boldsymbol{\Delta}_s)^{\dagger}$ is invertible under the assumptions that $\pi_i = E(\delta_i)$ is greater than zero for $i = 1, 2, \ldots, N$ and that $\boldsymbol{\Sigma}$ is positive definite. Thus, (2.3) is well defined.

c) By taking the expectation of (2.3), one immediately sees that $\hat{\theta}_{\text{GIP}}(\boldsymbol{\Sigma})$ is unbiased for estimating $\theta = \mathbf{y}'\mathbf{1}_{N\times 1}$. This is true for any positive definite $\boldsymbol{\Sigma}$. A poor choice of $\boldsymbol{\Sigma}$ may mean an estimator with a high variance, but it does not cause a bias.

d) Often, there is no closed-form formula for $E(\boldsymbol{\Delta}_s\boldsymbol{\Sigma}\boldsymbol{\Delta}_s)^{\dagger}$, but for single stage sampling plans at least, it can be easily approximated. One simply takes the average of a large number of values of $(\boldsymbol{\Delta}_s\boldsymbol{\Sigma}\boldsymbol{\Delta}_s)^{\dagger}$, each computed for a different sample obtained with the sampling plan. The computation does not require the knowledge of any of the variables of interest. It is a "desk exercise" in the sense that it does not require contacting the units. It can even be carried out before the actual sample is selected.

e) It is well known that for a total estimator utilizing a regression vector $\boldsymbol{\beta}$, $\hat{T}(\boldsymbol{\beta})$, is asymptotically equivalent in terms of bias and variance to the estimator $\hat{T}(\hat{\boldsymbol{\beta}}_s)$ where $\hat{\boldsymbol{\beta}}_s$ is an estimator that converges in probability to $\boldsymbol{\beta}$. Similarly, $\hat{\theta}_{\text{GIP}}(\hat{\boldsymbol{\Sigma}}_s)$ has the same asymptotic bias and variance as $\hat{\theta}_{\text{GIP}}(\boldsymbol{\Sigma})$ if the positive definite matrix $\hat{\boldsymbol{\Sigma}}_s$ converges in probability to the positive definite matrix $\boldsymbol{\Sigma}$. In essence, if the sample size is sufficiently large, the error introduced by estimating $\boldsymbol{\Sigma}$ by $\hat{\boldsymbol{\Sigma}}_s$ is negligible compared to the error in $\hat{\theta}_{\text{GIP}}(\boldsymbol{\Sigma})$ due to the sampling of units. All asymptotic results in this paper assume that the sampling plan is non informative (see, for example Cassel, Särndal and Wretman, 1977).

f) When $\mathbf{\Sigma} = \mathbf{I}$, then $\hat{\theta}_{\text{GIP}}(\mathbf{\Sigma})$ reduces to the ordinary inverse probability estimator, $\hat{\theta}_{\text{IP}}$, as given in (2.1). This is the justification for referring to $\hat{\theta}_{\text{GIP}}(\mathbf{\Sigma})$ as the generalized inverse probability estimator or the generalized Horvitz-Thompson estimator. It will be seen later, why this particular unbiased extension of the ordinary Horvitz-Thompson estimator is of interest.

g) An arbitrary symmetric positive definite matrix $\mathbf{\Sigma}$ may contain up to $N(N+1)/2$ distinct parameters. It is not feasible to specify so many values. If the sample s is utilized to estimate those parameters, the task of estimating $N(N+1)/2$ parameters from $n < N$ observations is clearly impossible. A simpler choice must be used. The simplest choice utilizes $\mathbf{\Sigma} = \mathbf{I}$, as seen in f). There are other choices that have a reasonable number of parameters. One example is given in Section 6.

h) For estimating a domain total $\mathbf{y}'\mathbf{c}$ where $\mathbf{c} = (c_1, \ldots, c_i, \ldots, c_N)'$ is a vector of known constants with $c_i = 1$ or 0 depending on whether unit i is in the domain or not, it suffices to replace (2.3), which is for estimating the population total, with $\mathbf{y}'(\mathbf{\Delta}_s\mathbf{\Sigma}\mathbf{\Delta}_s)^{\dagger}\left(E(\mathbf{\Delta}_s\mathbf{\Sigma}\mathbf{\Delta}_s)^{\dagger}\right)^{-1}\mathbf{c}$. The weight vector $(\mathbf{\Delta}_s\mathbf{\Sigma}\mathbf{\Delta}_s)^{\dagger}\left(E(\mathbf{\Delta}_s\mathbf{\Sigma}\mathbf{\Delta}_s)^{\dagger}\right)^{-1}\mathbf{c}$ varies with each domain described by $\mathbf{c}$; however the weight matrix, $(\mathbf{\Delta}_s\mathbf{\Sigma}\mathbf{\Delta}_s)^{\dagger}\left(E(\mathbf{\Delta}_s\mathbf{\Sigma}\mathbf{\Delta}_s)^{\dagger}\right)^{-1}$, does not depend on the domain. There are $N-n$ rows of this matrix that are nil. Even though there are potentially nN elements of the weight matrix that are non zero, post-multiplication by $\mathbf{c}$ will give the weight vector for any domain described by $\mathbf{c}$.

i) One possibility for the matrix $\mathbf{\Sigma}$ is one where all the diagonal elements are the same, and all the off-diagonal elements are the same. In this way, all the units are the same with respect to $\mathbf{\Sigma}$. However, if all units are the same with respect to the sampling plan, for example simple random sampling or Bernoulli sampling, and if all units are the same with respect to the parameter estimated, for example a total or an average for all units, then by symmetry, every sampled unit will have the same weight. Since both $\hat{\theta}_{\text{IP}}$ and $\hat{\theta}_{\text{GIP}}(\mathbf{\Sigma})$ are unbiased, both estimators will have the same weights. Nonetheless, for domain parameters, because some units are in the domain and some not, the symmetry argument no longer holds and the value of the off-diagonal elements of $\mathbf{\Sigma}$ may make a difference in $\hat{\theta}_{\text{GIP}}(\mathbf{\Sigma})$.

j) By setting $\mathbf{y} = \mathbf{1}_{N\times 1}$ in $\hat{\theta}_{\text{GIP}}(\mathbf{\Sigma})$, the estimator simply becomes the sum of all the weights of the sampled units and the parameter estimated becomes $\mathbf{1}_{1\times N}\mathbf{1}_{N\times 1} = N$, the known total number of units. However, the sum of the weights does not necessarily equal N. This does not bode well for the variance of $\hat{\theta}_{\text{GIP}}(\mathbf{\Sigma})$. To fix this, calibration can be used. Calibration was introduced by Deville and Särndal (1992). At its simplest, it would consist of scaling the inverse probability weights, generalized or not, by a common factor so that the resulting final weights do add up to N. Even for the ordinary inverse probability estimator, for some sampling plans, the sum of the design weights does not necessarily equal N, and here too, the solution lies in calibration. The subject of calibration is examined in the next section.

3. The generalized calibration estimator

The sum of the weights of an estimator is an estimate of the known population size, N. When the sampling plan is such that the sample size is not fixed, the ordinary inverse probability estimator of the population size will have a variance greater than zero. The sum of the weights of $\hat{\theta}_{\text{GIP}}(\boldsymbol{\Sigma})$, noted $S(\boldsymbol{\Sigma})$, is often a worse estimator of the population size than the sum of the weights of $\hat{\theta}_{\text{IP}}$; it will often vary, even when the sample size is fixed. An estimator whose estimates of the population size vary, cannot be seen as very reliable.

To fix the problem that the ordinary inverse probability estimator experiences when the sample size is variable, calibration can be used. The weights of $\hat{\theta}_{\text{CAL}}$ are calibrated so that their sum equals the population size, N. A similar fix can be made to the generalized estimator: $\hat{\theta}_{\text{GCAL}}(\boldsymbol{\Sigma}) = \left(N / S(\boldsymbol{\Sigma})\right)\hat{\theta}_{\text{GIP}}(\boldsymbol{\Sigma})$. The definition of $\hat{\theta}_{\text{GCAL}}(\boldsymbol{\Sigma})$ will be expanded to include the possibility of more calibration equations involving more auxiliary variables. The use of calibration equations was presented in Deville and Särndal (1992).

With an auxiliary variable matrix $\mathbf{X} \in \mathbb{R}^{N \times q}$ assumed to be of full rank and noting $\|\mathbf{v}\|_{\mathbf{M}} = (\mathbf{v}'\mathbf{M}\mathbf{v})^{1/2}$ the weighted Euclidean norm of the vector $\mathbf{v}$, the following problem is addressed:

Calibration Problem: Among the weight vectors $\mathbf{w}_s \in \mathbb{R}^N$ in the range of $\boldsymbol{\Delta}_s$, i.e., non-sampled units should have a weight of 0, which minimize $\left\|\mathbf{X}'\mathbf{w}_s - \mathbf{X}'\mathbf{1}_{N\times 1}\right\|_{\mathbf{T}}$, i.e., which "best" satisfy the q calibration equations, seek one that minimizes $\left\|\mathbf{w}_s - \mathbf{w}_{s\,\text{GIP}}(\boldsymbol{\Sigma})\right\|_{\mathbf{U}}$, i.e., as close as possible to the weights of $\hat{\theta}_{\text{GIP}}(\boldsymbol{\Sigma})$, where $\mathbf{T} \in \mathbb{R}^{q \times q}$ and $\mathbf{U} \in \mathbb{R}^{N \times N}$ are positive definite matrices.

Weights, $\mathbf{w}_s$, that satisfy the calibration equations, $\mathbf{X}'\mathbf{w}_s = \mathbf{X}'\mathbf{1}_{N\times 1}$, do not always exist, especially if the number of equations, q, is high relative to the sample size. To prepare for this eventuality, the matrix $\mathbf{T}$ is at the statistician's disposal for specifying the relative importance of the q calibration equations. The matrix $\mathbf{U}$ specifies the relative importance given to each unit when measuring the distance from $\mathbf{w}_{s\,\text{GIP}}(\boldsymbol{\Sigma})$. This formulation of the calibration problem generalizes that of Théberge (1999), where $\mathbf{T}$ and $\mathbf{U}$ were diagonal matrices, and the inverse probability, or Horvitz-Thompson, weights were used instead of the generalized inverse probability weights.

The solution to the calibration problem yields

$$\begin{aligned}\hat{\theta}_{\text{GCAL}}(\boldsymbol{\Sigma}) &= \mathbf{y}'\mathbf{w}_{s\,\text{GCAL}}(\boldsymbol{\Sigma}) \\ &= \hat{\mathbf{y}}'\mathbf{1}_{N\times 1} + (\mathbf{y} - \hat{\mathbf{y}})'\,\mathbf{w}_{s\,\text{GIP}}(\boldsymbol{\Sigma}),\end{aligned} \tag{3.1}$$

where $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ with

$$\hat{\boldsymbol{\beta}} = \mathbf{T}^{1/2}\left(\mathbf{T}^{1/2}\mathbf{X}'(\boldsymbol{\Delta}_s\mathbf{U}\boldsymbol{\Delta}_s)^{\dagger}\mathbf{X}\mathbf{T}^{1/2}\right)^{\dagger}\mathbf{T}^{1/2}\mathbf{X}'(\boldsymbol{\Delta}_s\mathbf{U}\boldsymbol{\Delta}_s)^{\dagger}\,\mathbf{y}. \tag{3.2}$$

The estimator $\hat{\theta}_{\text{GCAL}}(\boldsymbol{\Sigma})$ is asymptotically unbiased. Also, if $\hat{\boldsymbol{\Sigma}}_s \to \boldsymbol{\Sigma}$ in probability, then the bias and variance of $\hat{\theta}_{\text{GCAL}}(\hat{\boldsymbol{\Sigma}}_s)$ are asymptotically the same as those of $\hat{\theta}_{\text{GCAL}}(\boldsymbol{\Sigma})$. The rate at which $\hat{\boldsymbol{\Sigma}}_s \to \boldsymbol{\Sigma}$ will depend on the estimator $\hat{\boldsymbol{\Sigma}}_s$ and on the number of parameters in $\boldsymbol{\Sigma}$.

The difference between $\hat{\theta}_{\text{GCAL}}(\boldsymbol{\Sigma})$ and the ordinary calibration estimator, $\hat{\theta}_{\text{GCAL}}(\mathbf{I}_{N\times N})$, is simply the use of generalized inverse probability weights to estimate the sum of the residues, rather than the usual inverse probability weights. This was to be expected given that in one case we are, in the calibration problem, seeking weights that minimize $\left\|\mathbf{w}_s - \mathbf{w}_{s\,\text{GIP}}(\boldsymbol{\Sigma})\right\|_{\mathbf{U}}$, instead of weights that minimize $\left\|\mathbf{w}_s - \mathbf{w}_{s\,\text{IP}}\right\|_{\mathbf{U}}$, where $\mathbf{w}_{s\,\text{IP}} = \mathbf{w}_{s\,\text{GIP}}(\mathbf{I}_{N\times N})$ are the usual inverse probability weights.

The following result is proven in the Appendix: for any $\boldsymbol{\alpha} \in \mathbb{R}^N$, if $\boldsymbol{\Delta}_s\boldsymbol{\alpha}$ is in the range of $\boldsymbol{\Delta}_s\mathbf{X}$, then the weighted sum of residuals, $(\mathbf{y}-\hat{\mathbf{y}})'(\boldsymbol{\Delta}_s\mathbf{U}\boldsymbol{\Delta}_s)^{\dagger}\boldsymbol{\alpha}$, is zero. A vector $\mathbf{v}$ is said to be in the range of a matrix $\mathbf{F}$ is there exists a vector $\boldsymbol{\lambda}$ such that $\mathbf{v} = \mathbf{F}\boldsymbol{\lambda}$. In particular, if the matrix $\mathbf{U}$ is diagonal and written $\mathbf{U} = \mathbf{A}^{-1}\mathbf{D}$, where $\mathbf{A} = \left(E(\boldsymbol{\Delta}_s)\right)^{-1}$ is the diagonal matrix of the ordinary inverse probability weights and $\mathbf{D} \in \mathbb{R}^{N\times N}$ is an arbitrary positive diagonal matrix, then with $\boldsymbol{\alpha} = \mathbf{D}\mathbf{c}$ the result gives that $(\mathbf{y}-\hat{\mathbf{y}})'\left(\boldsymbol{\Delta}_s\mathbf{A}^{-1}\mathbf{D}\boldsymbol{\Delta}_s\right)^{\dagger}\mathbf{D}\mathbf{c} = (\mathbf{y}-\hat{\mathbf{y}})'\mathbf{A}\boldsymbol{\Delta}_s\mathbf{c}$ is zero if $\boldsymbol{\Delta}_s\mathbf{D}\mathbf{c}$ is in the range of $\boldsymbol{\Delta}_s\mathbf{X}$. This is similar to result 6.5.1 of Särndal, Swensson and Wretman (1992), for example, where $\mathbf{c}$ is a vector of ones and the diagonal elements of $\mathbf{D}$ are variances.

It can be seen from the form of (3.1), that $\hat{\theta}_{\text{GCAL}}(\boldsymbol{\Sigma})$ is also a regression estimator that uses a model ξ such that $E_{\xi}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$. Despite the notation used in (3.2), calibration estimators do not use models, instead there are calibration equations. When viewed as a regression estimator, it is important to realize that $\hat{\theta}_{\text{GCAL}}(\boldsymbol{\Sigma})$ is asymptotically design unbiased, regardless of the choice of the model parameter $\boldsymbol{\beta}$, and regardless of the choice of the positive definite matrix $\boldsymbol{\Sigma}$.

4. The choice of the positive definite matrix $\boldsymbol{\Sigma}$

Different choices for $\boldsymbol{\Sigma}$ will generally lead to different generalized inverse probability estimators and different generalized calibration estimators. The advantage of the generalization of the inverse probability estimator comes from its use in a generalization of calibration, as seen in Section 3, and the optimality of generalized calibration, as discussed in Section 5. It will be seen that a matrix $\boldsymbol{\Sigma}$ is an appropriate choice to use for $\hat{\theta}_{\text{GIP}}(\boldsymbol{\Sigma})$, if a model ξ with $V_{\xi}(\mathbf{y}) = \boldsymbol{\Sigma}$ is an appropriate model for $\mathbf{y}$. Even if the assumption that $V_{\xi}(\mathbf{y}) = \boldsymbol{\Sigma}$ is wrong, the estimator $\hat{\theta}_{\text{GIP}}(\boldsymbol{\Sigma})$ remains design unbiased and the estimator $\hat{\theta}_{\text{GCAL}}(\boldsymbol{\Sigma})$ remains asymptotically design unbiased. The generalized calibration estimators with $\boldsymbol{\Sigma} = V_{\xi}(\mathbf{y})$ can be said to be model assisted as opposed to model based or model dependent (see Särndal et al., 1992, Section 6.7). The ordinary calibration estimators, $\hat{\theta}_{\text{CAL}}$ use (3.1) with $\boldsymbol{\Sigma} = \mathbf{I}$. A model that fits the population perfectly is not necessary, but hopefully a better model than one with $V_{\xi}(\mathbf{y}) = \mathbf{I}$ can be utilized. In fact, if $\boldsymbol{\Sigma}$ is any positive diagonal matrix, then $\hat{\theta}_{\text{GIP}}(\boldsymbol{\Sigma})$ will result in the ordinary inverse probability estimator, and the generalized calibration estimator will result in the ordinary calibration estimator. Often, a more appropriate model for $\mathbf{y}$ would have $V_{\xi}(\mathbf{y})$ non-diagonal. As for the variance of $\hat{\theta}_{\text{GIP}}(\boldsymbol{\Sigma})$, it may be higher than that of the ordinary inverse probability estimator, even if $V_{\xi}(\mathbf{y}) = \boldsymbol{\Sigma}$. It is the calibration of $\hat{\theta}_{\text{GIP}}(\boldsymbol{\Sigma})$ that yields, as will be seen in Section 5, an optimal estimator.

The use of a block-diagonal matrix simplifies the computation of inverses needed in (2.3). Blocks may correspond to persons of a household, students of a class, workers of an establishment, dwellings of a

block, etc. It is often natural for units belonging to the same block to have a correlated variable of interest. For example, how one worker rates their employer is likely correlated with the rating of another worker of the same employer; the race or religion of a couple is often the same. In such cases, a multistage sampling plan would often be used, but it will be assumed here that a single stage plan is used. This could be because a single stage sampling plan was more suitable for other variables of interest of the same survey, or because some unit level characteristics are so important, that it is desirable to stratify at the population level so that the sample can be targeted at certain strata. For example, it may be important to stratify persons by age, but households can't be stratified by age.

In the simulation presented in this paper, the vaccination status of individuals in two-person households is made to be correlated. An extreme case presents itself if the blocks are persons of a same household and the variable of interest is household income. In such a case the correlation is perfect, and lines of $\mathbf{\Sigma}$ corresponding to persons from a same household should be identical. Such a matrix $\mathbf{\Sigma}$ is not positive definite, but it is the limit of a sequence of positive definite matrices, and the limit of the corresponding sequence of generalized inverse probability estimators could be computed. The example (1.2) given in the introduction is based on this idea.

If $\mathbf{\Sigma}$ is block-diagonal with blocks $\mathbf{\Sigma}_1, \mathbf{\Sigma}_2, \ldots, \mathbf{\Sigma}_B$, then because both the Moore-Penrose inverse and the ordinary inverse of a block-diagonal matrix is the block-diagonal matrix of inverses, the estimator $\hat{\theta}_{\text{GIP}}(\mathbf{\Sigma})$ can be decomposed into

$$\begin{aligned}\hat{\theta}_{\text{GIP}}(\mathbf{\Sigma}) &= \sum_{b=1}^{B} \hat{\theta}_{\text{GIP}\,b}(\mathbf{\Sigma}_b) \\ &= \sum_{b=1}^{B} \mathbf{y}_b' \left(\mathbf{\Delta}_{s_b}\mathbf{\Sigma}_b\mathbf{\Delta}_{s_b}\right)^{\dagger} \left(E\left(\mathbf{\Delta}_{s_b}\mathbf{\Sigma}_b\mathbf{\Delta}_{s_b}\right)^{\dagger}\right)^{-1} \mathbf{1}_{N_b \times 1}\,, \end{aligned} \tag{4.1}$$

where N_b is the size of block b, $\mathbf{y}_b$ and $\mathbf{\Delta}_{s_b}$ are the sub-vector and sub-matrix respectively, which correspond to block b.

If the population is partitioned into blocks of correlated units, the variable defining the blocks must be on the frame. But that variable need not be perfect. For example, a unit's household may only be known at the time of the survey, but using an outdated household variable available on the frame will still be useful, while not introducing any bias. It simply means that the strength borrowed by the generalized inverse probability estimator from the correlations will be reduced. On the other hand, the strength borrowed from the correlations by the ordinary inverse probability estimator is nil.

If a positive definite estimator $\hat{\mathbf{\Sigma}}_s$ converges to a positive definite $\mathbf{\Sigma}$ in probability, then the bias and variance of $\hat{\theta}_{\text{GIP}}(\hat{\mathbf{\Sigma}}_s)$ are asymptotically the same as those of $\hat{\theta}_{\text{GIP}}(\mathbf{\Sigma})$. In practice, even if the general form of $\mathbf{\Sigma}$ depends on $N(N-1)/2$ covariances, the number of parameters in $\mathbf{\Sigma}$ should be small compared to the sample size. Using the inverse probability estimator means assuming all covariances are zero. When using the generalized inverse probability estimator, one could assume that those covariances depend on a few parameters, and that those parameters are considered fixed, rather than estimated from the sample. In the examples of Section 6, $\mathbf{\Sigma}$ depends on only one parameter, ρ, and its value is assumed to be 1.

5. The generalized Godambe-Joshi lower bound

For any unbiased estimator $\hat{\theta}$ of the population total θ, if $V_p\left(\hat{\theta}\right)$ is the variance of $\hat{\theta}$ under the sampling plan, Godambe and Joshi (1965) have given a lower bound for the value of $E_\xi V_p\left(\hat{\theta}\right)$ under the assumption that the variance matrix $V_\xi(\mathbf{y})$ was diagonal. That lower bound is the sum of the elements of the diagonal matrix $\left(\left(E\left(\mathbf{\Delta}_s\right)\right)^{-1}-\mathbf{I}\right)V_\xi(\mathbf{y})$. That result is generalized in the following paragraph.

For any linear unbiased total estimator, $\hat{\theta}$, if $V_\xi(\mathbf{y})$ is positive definite, then $E_\xi V_p\left(\hat{\theta}\right)$ is not lower than the sum of the elements of the matrix $\left(E\left(\mathbf{\Delta}_s V_\xi(\mathbf{y})\mathbf{\Delta}_s\right)^{\dagger}\right)^{-1}-V_\xi(\mathbf{y})$. It is easily verified that the usual Godambe-Joshi lower bound is obtained if $V_\xi(\mathbf{y})$ is diagonal.

Just as the calibration estimator asymptotically attains the Godambe-Joshi lower bound, the generalized calibration estimator with $\mathbf{\Sigma}=V_\xi(\mathbf{y})$, asymptotically attains the generalized Godambe-Joshi lower bound, regardless of the value of the matrices $\mathbf{X}$, $\mathbf{T}$ and $\mathbf{U}$. The link between the value of those three matrices and the value of $V_\xi(\mathbf{y})$ is not examined in this paper, but the calibration problem stated in Section 3 does clarify the role of each of those matrices. The derivation of the generalized lower bound and the proof of the optimality of the generalized calibration estimator are given in Théberge (2017).

The fact that $\hat{\theta}_{\text{GCAL}}\left(V_\xi(\mathbf{y})\right)$ asymptotically attains the generalized Godambe-Joshi lower bound shows that the generalized inverse probability estimator performs well when applied to residuals, as it does in (3.1), even though it is not recommended in general. Similarly, the ordinary inverse probability estimator can be inefficient if the sample size is random, but will perform well if applied to residuals.

It should be noted that, contrary to the ordinary Godambe-Joshi lower bound, the generalized lower bound applies only to *linear* unbiased estimators. In fact, an example with $V_\xi(\mathbf{y})$ not diagonal, of a non-linear unbiased estimator which does better than the lower bound is given in Théberge (2017).

6. Examples

There are cases simple enough for $\hat{\theta}_{\text{GIP}}(\mathbf{\Sigma})$ to be given explicitly. Say $\mathbf{\Sigma}(\rho)$ is a block-diagonal matrix where each of $N_p=N/2$ blocks equals $\sigma^2\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}$, with $-1<\rho<1$. Such a block-diagonal matrix corresponds to a model of a population which can be partitioned into pairs $\{2i-1, 2i\}$ $(i=1,2,\ldots,N_p)$ where, within a pair, the variable of interest is correlated. Then, (2.3) reduces to

$$\hat{\theta}_{\text{GIP}}\left(\mathbf{\Sigma}(\rho)\right) = \sum_{i=1}^{N_p}\frac{a_{2i-1}y_{2i-1}+a_{2i}y_{2i}}{\left(\pi_{2i-1}\pi_{2i}\left(1-\rho^2\right)+\left(\pi_{2i-1}+\pi_{2i}-\pi_{2i-12i}\right)\pi_{2i-12i}\rho^2\right)}, \tag{6.1}$$

where

$$\begin{aligned} a_{2i-1} &= \delta_{2i-1}\left[\pi_{2i}\left(1-\rho^2\right)+\pi_{2i-12i}\rho\left(1+\rho\right)\right]+\delta_{2i-1}\delta_{2i}\left[\rho^2\pi_{2i}-\rho\pi_{2i-1}-\rho^2\pi_{2i-12i}\right]\\ a_{2i} &= \delta_{2i}\left[\pi_{2i-1}\left(1-\rho^2\right)+\pi_{2i-12i}\rho\left(1+\rho\right)\right]+\delta_{2i-1}\delta_{2i}\left[\rho^2\pi_{2i-1}-\rho\pi_{2i}-\rho^2\pi_{2i-12i}\right]. \end{aligned} \tag{6.2}$$

Once again, this generalized inverse probability estimator is unbiased, for any value of ρ, "correct" or not. It is seen that as expected, when $\rho=0$, the estimator reduces to the inverse probability estimator. The

value of ρ cannot simply be set to one in (6.1), because $\boldsymbol{\Sigma}(1)$ is not positive definite. However, the limit of (6.1) as $\rho \to 1$ results in the estimator (1.2) given in the Introduction, $\hat{\theta}_{\text{LIM}}$. It can be calibrated so that the sum of the weights is equal to N. If the probabilities of inclusion do not vary with $i = 1, 2, \ldots, N_p$, the resulting estimator is

$$\hat{\theta}_{\text{LCAL}} = \frac{N_p}{\nu_p} \sum_{i=1}^{N_p} 2y_{2i-1}\delta_{2i-1} + 2y_{2i}\delta_{2i} - \left(y_{2i-1} + y_{2i}\right)\delta_{2i-1}\delta_{2i} + \left(y_{2i-1} - y_{2i}\right)\delta_{2i-1}\delta_{2i}\pi_{\text{diff } i}, \qquad (6.3)$$

where $\nu_p = \sum_{i=1}^{N_p} \left(\delta_{2i-1} + \delta_{2i} - \delta_{2i-1}\delta_{2i}\right)$ is the number of pairs with at least one unit in the sample. It is easy to verify, by setting $\mathbf{y} = \mathbf{1}_{N\times 1}$ in (6.3), that the sum of the weights of $\hat{\theta}_{\text{LCAL}}$ is equal to $2N_p = N$. The generalized calibration estimator (6.3) is optimized for $\rho \to 1$, but it can still have a lower variance than both, the inverse probability estimator and the ordinary calibration estimator, if the correlation between the units of a pair is strong (for example, race, religion or education level of a couple). Since a variable indicating which unit is paired with which, must be on the frame, a calibration at the pair level would be possible. The calibration would ensure that the sum of the weights of the sampled units of a pair would equal 2. However, the low number of observations per calibration group would not ensure the validity of asymptotic results and could result in significant biases.

There are modified versions of the generalized inverse probability estimator and of the generalized calibration estimator. The modified versions have the advantage of having a closed form; there is no need to compute the expectation of $\left(\boldsymbol{\Delta}_s\boldsymbol{\Sigma}\boldsymbol{\Delta}_s\right)^{\dagger}$. They also do not rely on the Moore-Penrose inverse. For a positive definite matrix $\boldsymbol{\Sigma}$, they are defined as

$$\hat{\theta}_{\text{MGIP}}\left(\boldsymbol{\Sigma}\right) = \mathbf{y}'\boldsymbol{\Delta}_s\boldsymbol{\Sigma}^{-1}\boldsymbol{\Delta}_s\left(\boldsymbol{\Sigma}^{-1} \circ \boldsymbol{\Pi}\right)^{-1}\mathbf{1}_{N\times 1} \qquad (6.4)$$

and

$$\hat{\theta}_{\text{MGCAL}}\left(\boldsymbol{\Sigma}\right) = \hat{\mathbf{y}}'\mathbf{1}_{N\times 1} + \left(\mathbf{y} - \hat{\mathbf{y}}\right)'\mathbf{w}_{s\,\text{MGHT}}\left(\boldsymbol{\Sigma}\right), \qquad (6.5)$$

where $\boldsymbol{\Pi} = \left(\pi_{kl}\right) = \left(E\left(\delta_k\delta_l\right)\right) \in \mathbb{R}^{N\times N}$ is the matrix of second order probabilities of inclusion, $\mathbf{w}_{s\,\text{MGIP}}\left(\boldsymbol{\Sigma}\right)$ is the vector of weights of $\hat{\theta}_{\text{MGIP}}\left(\boldsymbol{\Sigma}\right)$, and $\circ$ denotes the Hadamard product, i.e., element-wise multiplication. With $\hat{\theta}_{\text{MGIP}}\left(\boldsymbol{\Sigma}\right)$, the "probability" part of the phrase "inverse probability" is $\boldsymbol{\Sigma}^{-1} \circ \boldsymbol{\Pi}$. The modified generalized estimators are also unbiased, or at least asymptotically unbiased in the case of $\hat{\theta}_{\text{MGCAL}}\left(\boldsymbol{\Sigma}\right)$. The usual estimators $\hat{\theta}_{\text{IP}}$ and $\hat{\theta}_{\text{CAL}}$ are obtained if $\boldsymbol{\Sigma} = \mathbf{I}$.

If $\rho \to 1$, the modified generalized inverse probability estimator, $\hat{\theta}_{\text{MGIP}}\left(\boldsymbol{\Sigma}\left(\rho\right)\right)$, becomes:

$$\hat{\theta}_{\text{MLIM}} = \sum_{i=1}^{N_p} w_{2i-1}y_{2i-1} + w_{2i}y_{2i}, \qquad (6.6)$$

where

$$w_{2i-1} = \frac{\delta_{2i-1}\left(\pi_{2i} + \pi_{2i-1\;2i}\right) - \delta_{2i-1}\delta_{2i}\left(\pi_{2i-1} + \pi_{2i-1\;2i}\right)}{\pi_{2i-1}\pi_{2i} - \pi_{2i-1\;2i}^2} \qquad (6.7)$$

and

$$w_{2i} = \frac{\delta_{2i}\left(\pi_{2i-1} + \pi_{2i-1\ 2i}\right) - \delta_{2i-1}\delta_{2i}\left(\pi_{2i} + \pi_{2i-1\ 2i}\right)}{\pi_{2i-1}\pi_{2i} - \pi_{2i-1\ 2i}^2}. \tag{6.8}$$

If the sampling plan is such that $\pi_{2i} = \pi_{2i-1}$ for any $i = 1, 2, \ldots, N_p$, and if both units of that pair are sampled, then the weights of both units will be zero. That some sampled units may not contribute to the estimator, in some circumstances, is an undesirable property of $\hat{\theta}_{\text{MLIM}}$.

One characteristic of the estimator $\hat{\theta}_{\text{LIM}}$ is somewhat surprising. It is constructed in such a way that for each observed pair, that is each pair with at least one unit in the sample, the numerator in (1.2) corresponds to a value for the pair's variable of interest total. The numerator of the ith term is 0 if neither unit $2i-1$ nor unit $2i$ are observed, it is $2y_{2i-1}$ if only unit $2i-1$ of the pair is sampled, it is $2y_{2i}$ if only unit $2i$ of the pair is sampled, and it is $\left(y_{2i-1} + y_{2i}\right) + \left(y_{2i-1} - y_{2i}\right)\pi_{\text{diff } i}$ if both units of the pair are sampled. The unexpected characteristic is that when both units of a pair i $\left(i = 1, 2, \ldots, N_p\right)$ are observed, the estimated value for the pair's total is not the known total $y_{2i-1} + y_{2i}$. This is the motivation for yet another estimator and its calibrated version, where the estimate for a pair, while still being unbiased, will agree with the known total when both units of the pair are sampled. The alternative estimators are

$$\hat{\theta}_{\text{ALIM}} = \sum_{i=1}^{N_p} \frac{\left(a_i\delta_{2i-1} + b_i\delta_{2i-1}\delta_{2i}\right)y_{2i-1} + \left(c_i\delta_{2i} + d_i\delta_{2i-1}\delta_{2i}\right)y_{2i}}{\pi_{2i-1} + \pi_{2i} - \pi_{2i-1\,2i}} \tag{6.9}$$

and

$$\hat{\theta}_{\text{ALCAL}} = \hat{\mathbf{y}}'\mathbf{1}_{N\times 1} + \left(\mathbf{y} - \hat{\mathbf{y}}\right)'\mathbf{w}_{s\,\text{ALIM}}\left(\boldsymbol{\Sigma}\right), \tag{6.10}$$

where $\mathbf{w}_{s\,\text{ALIM}}$ is the vector of weights of $\hat{\theta}_{\text{ALIM}}$, $a_i + b_i = c_i + d_i = 1$, motivated by what is wanted when both units of the pair are sampled, and in order to have $\hat{\theta}_{\text{ALIM}}$ unbiased, one should have $a_i\pi_{2i-1} + b_i\pi_{2i-1\,2i} = c_i\pi_{2i} + d_i\pi_{2i-1\,2i} = \pi_{2i-1} + \pi_{2i} - \pi_{2i-1\,2i}$. Therefore, for $i = 1, 2, \ldots, N_p$,

$$\begin{aligned}
a_i &= \frac{\pi_{2i-1} + \pi_{2i} - 2\pi_{2i-1\,2i}}{\pi_{2i-1} - \pi_{2i-1\,2i}} \\
b_i &= \frac{\pi_{2i-1\,2i} - \pi_{2i}}{\pi_{2i-1} - \pi_{2i-1\,2i}} \\
c_i &= \frac{\pi_{2i-1} + \pi_{2i} - 2\pi_{2i-1\,2i}}{\pi_{2i} - \pi_{2i-1\,2i}} \\
d_i &= \frac{\pi_{2i-1\,2i} - \pi_{2i-1}}{\pi_{2i} - \pi_{2i-1\,2i}}.
\end{aligned} \tag{6.11}$$

7. Simulation results

In this simulation, estimators from the preceding section will be compared to the ordinary inverse probability estimator and the ordinary calibrated estimator. A population of 2,000 individuals grouped into 1,000 two-person households was generated. Persons $2i$ and $2i-1$ for $i = 1, 2, \ldots, 1{,}000$ belong to the same household. A variable of interest y takes the value 1 to represent a vaccinated person, and it takes

the value 0 to represent an unvaccinated person. To simulate how vaccination status can be correlated within household, the method of Lunn and Davies (1998) was used to generate pairs of correlated Bernoulli variables with a probability of 0.7 of a value of 1 and a correlation of 0.8. The actual population generated has 254 households where neither person is vaccinated, 660 households where both are vaccinated, 44 households where only the person with an odd label is vaccinated, and 42 households where only the person with an even label is vaccinated. The total number of persons vaccinated is $660 \times 2 + 44 + 42 = 1{,}406$ for a vaccination rate of 0.703. The correlation between persons of the same household is $\left(0.66 - 0.704 \times 0.702\right) \Big/ \left(\sqrt{0.704 \times 0.296} \times \sqrt{0.702 \times 0.298}\right) = 0.7941$.

The population was sampled 10,000 times. Each household $i\left(i = 1, 2, \ldots, N_p\right)$ was sampled independently; the probability of selecting both units was 0.05, the probability of selecting only unit $2i-1$ was 0.10, and the probability of selecting only unit $2i$ was 0.05. Thus, for each sample, the probabilities of inclusion were $\pi_{2i-1} = 0.15$, $\pi_{2i} = 0.1$ and $\pi_{2i-1\,2i} = 0.05$. This means $\pi_{\text{diff } i} = \left(\pi_{2i} - \pi_{2i-1}\right) \Big/ \pi_{2i-1\,2i}$ in (1.2) was chosen to not be zero. This is because when $\pi_{\text{diff } i}$ is zero, $\hat{\theta}_{\text{LIM}}$ is a somewhat obvious choice: it is an inverse probability estimator based on pairs where the pair is given a value of $2y_{2i-1}$ if only unit $2i-1$ is sampled, a value of $2y_{2i}$ if only unit $2i$ is sampled, and a value of $y_{2i-1} + y_{2i}$ if both units are sampled. Combined with calibration, it is an obvious competitor to the ordinary calibration estimator. Why not base the estimator on pairs in this way, rather than units, if there is a strong correlation between units of a pair? When $\pi_{\text{diff } i}$ is zero, it is also true that $\hat{\theta}_{\text{LIM}} = \hat{\theta}_{\text{ALIM}}$. It is interesting to find out how $\hat{\theta}_{\text{LIM}}$ compares when $\pi_{\text{diff } i}$ is not zero. For each sample, eight estimators of the total were calculated: the inverse probability estimator, the ordinary calibrated estimator, the generalization of the inverse probability estimator and its calibrated version, the modified generalized inverse probability estimator and its calibrated version, and finally the alternative estimator and its calibrated version. For the generalized and modified generalized estimators, including their calibrated versions, $\mathbf{\Sigma}(\rho)$ with $\rho \to 1$ was used, as explained in the examples of the preceding section. The simple closed-form formulae of that section can thus be used. For the calibration, $\mathbf{X} = \mathbf{1}_{N \times 1}$ with $\mathbf{T} = 1$ and $\mathbf{U} = \mathbf{I}_{N \times N}$ yields $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \frac{\sum_{k \in s} y_k}{\sum_{k \in s} x_k} \mathbf{1}_{N \times 1}$. The average total and the variance over the 10,000 repetitions are given in Table 7.1.

Table 7.1
Simulation results comparing eight estimators

Estimator and lower bounds	Total	Variance
Inverse probability	1,406.60	13,326
Calibrated inverse probability	1,407.38	3,856
Generalized inverse probability	1,406.41	11,226
Calibrated generalized inverse probability	1,407.08	3,419
Modified generalized inverse probability	1,406.37	16,337
Calibrated modified generalized inverse probability	1406.68	4,932
Alternative	1,406.61	12,447
Calibrated alternative	1,407.12	3,697
Generalized Godambe-Joshi lower bound ($\rho =$ 0.8)		3,408
Generalized Godambe-Joshi lower bound ($\rho \to 1$)		3,360

All eight estimators are either unbiased or asymptotically unbiased, so as expected, the observed bias of each estimator is negligible, since the real population total is 1,406.

The observed variances show that only the four calibrated estimators have reasonable variances. With the sampling plan used for this simulation, only the calibrated estimators can estimate the known population total with zero variance.

The calibrated generalized inverse probability estimator, with a variance of 3,419, performs best. This despite being calculated assuming that the correlation between the units of a pair is one. It should be remembered that the calibrated inverse probability estimator, with a variance of 3,856, is a special case of the calibrated generalized inverse probability estimator, but it is computed assuming that the correlation between the units of a pair is zero. The calibrated alternative estimator, which contrary to the other estimators, has been defined only for a household size of 2, has a variance somewhere in between that of the calibrated versions of the inverse probability and generalized inverse probability estimators. Finally, the calibrated modified generalized estimator had the highest variance of the four calibrated estimators.

The generalized Godambe-Joshi lower bound with the variance matrix, $V_\xi(\mathbf{y})$, of the model ξ used to generate $\mathbf{y}$ is 3,408. This is the asymptotic variance that could be expected of the calibrated generalized estimator, if it had been calculated with a matrix $\mathbf{\Sigma} = V_\xi(\mathbf{y})$ based on the correct model ξ, where the correlation between units of a pair is 0.8. If $\mathbf{\Sigma}(\rho)$ is defined as in the preceding section, and $\mathrm{GJ}(\mathbf{\Sigma}(\rho))$ is the generalized Godambe-Joshi lower bound for the positive definite variance matrix $V_\xi(\mathbf{y}) = \mathbf{\Sigma}(\rho)$, then the limit as $\rho \to 1$ of $\mathrm{GJ}(\mathbf{\Sigma}(\rho))$ is 3,360. This is the variance that could be expected of the generalized calibration estimator, if the correlation between units of a same pair was one.

8. Summary

The concept of inverse probability estimation can be generalized with a positive definite matrix $\mathbf{\Sigma}$. There is then a whole family of unbiased estimators parameterized by $\mathbf{\Sigma}$ where one member, with $\mathbf{\Sigma} = \mathbf{I}_{N \times N}$ is the usual inverse probability estimator. The concept of calibration can also be generalized so that weights close to those of the generalized inverse probability estimator are sought. The Godambe and Joshi lower bound of $E_\xi V_p(\hat{\theta})$ can also be generalized to a model ξ where the variance matrix $V_\xi(\mathbf{y})$ is not necessarily diagonal. The calibrated generalized inverse probability estimator, with $\mathbf{\Sigma} = V_\xi(\mathbf{y})$, asymptotically attains the generalized lower bound for any linear unbiased estimator $\hat{\theta}$. The new estimators are model assisted, not model based. They remain unbiased, or at least asymptotically unbiased, even if $\mathbf{\Sigma} \neq V_\xi(\mathbf{y})$.

Examples where the new estimators can be given an explicit form have been presented. Simulations comparing those new estimators with the usual ones have been done. Those simulations show that, while remaining asymptotically unbiased, significant improvements in variance can be obtained in situations where there is significant correlation between some units of the population, as for example there would be,

between persons of a same household with regards to vaccination status. Improvements in variance can still be made, even with $\mathbf{\Sigma} \neq V_{\xi}(\mathbf{y})$.

Acknowledgements

I would like to thank the Associate Editor and the referees for their constructive comments and suggestions to improve the paper.

Appendix

Proof that with $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\beta}} = \mathbf{T}^{1/2}\left(\mathbf{T}^{1/2}\mathbf{X}'\left(\mathbf{\Delta}_s\mathbf{U}\mathbf{\Delta}_s\right)^{\dagger}\mathbf{X}\mathbf{T}^{1/2}\right)^{\dagger}\mathbf{T}^{1/2}\mathbf{X}'\left(\mathbf{\Delta}_s\mathbf{U}\mathbf{\Delta}_s\right)^{\dagger}\mathbf{y}$, where $\mathbf{T}$ and $\mathbf{U}$ are positive definite, then for any $\boldsymbol{\alpha} \in \mathbb{R}^N$, if $\mathbf{\Delta}_s\boldsymbol{\alpha}$ is in the range of $\mathbf{\Delta}_s\mathbf{X}$, then the weighted sum of residuals, $\left(\mathbf{y} - \hat{\mathbf{y}}\right)'\left(\mathbf{\Delta}_s\mathbf{U}\mathbf{\Delta}_s\right)^{\dagger}\boldsymbol{\alpha}$, is zero.

First,

$$\begin{aligned}\left(\mathbf{y} - \hat{\mathbf{y}}\right)'\left(\mathbf{\Delta}_s\mathbf{U}\mathbf{\Delta}_s\right)^{\dagger}\boldsymbol{\alpha} &= \mathbf{y}'\left(\mathbf{\Delta}_s\mathbf{U}\mathbf{\Delta}_s\right)^{\dagger}\left[\mathbf{I} - \mathbf{X}\mathbf{T}^{1/2}\left(\mathbf{T}^{1/2}\mathbf{X}'\left(\mathbf{\Delta}_s\mathbf{U}\mathbf{\Delta}_s\right)^{\dagger}\mathbf{X}\mathbf{T}^{1/2}\right)^{\dagger}\mathbf{T}^{1/2}\mathbf{X}'\left(\mathbf{\Delta}_s\mathbf{U}\mathbf{\Delta}_s\right)^{\dagger}\right]\boldsymbol{\alpha} \\ &= \mathbf{y}'\mathbf{M}\boldsymbol{\alpha}. \end{aligned} \tag{A.1}$$

With $\mathbf{\Delta}_s$ being an orthogonal projection, note that by Lemma 2 of Théberge (2017), $\mathbf{M} = \mathbf{M}\mathbf{\Delta}_s$, and that by the properties of the Moore-Penrose inverse, $\mathbf{M}\mathbf{\Delta}_s\mathbf{X}\mathbf{T}^{1/2}\left(\mathbf{T}^{1/2}\mathbf{X}'\left(\mathbf{\Delta}_s\mathbf{U}\mathbf{\Delta}_s\right)^{\dagger}\mathbf{X}\mathbf{T}^{1/2}\right)^{\dagger} = \mathbf{0}$. For $\mathbf{T}$ and $\mathbf{U}$ of full rank, one has that the rank of $\mathbf{\Delta}_s\mathbf{X}\mathbf{T}^{1/2}\left(\mathbf{T}^{1/2}\mathbf{X}'\left(\Delta_s\mathbf{U}\mathbf{\Delta}_s\right)^{\dagger}\mathbf{X}\mathbf{T}^{1/2}\right)^{\dagger}$ equals the rank of $\mathbf{\Delta}_s\mathbf{X}\mathbf{T}^{1/2}$. It then follows that the range of $\mathbf{\Delta}_s\mathbf{X}$, which equals the range of $\mathbf{\Delta}_s\mathbf{X}\mathbf{T}^{1/2}$, equals the range of $\mathbf{\Delta}_s\mathbf{X}\mathbf{T}^{1/2}\left(\mathbf{T}^{1/2}\mathbf{X}'\left(\mathbf{\Delta}_s\mathbf{U}\mathbf{\Delta}_s\right)^{\dagger}\mathbf{X}\mathbf{T}^{1/2}\right)^{\dagger}$ by exercise 1.10 of Ben-Israel and Greville (2002). Therefore, if $\mathbf{\Delta}_s\boldsymbol{\alpha}$ is in the range of $\mathbf{\Delta}_s\mathbf{X}$ which equals the range of $\mathbf{\Delta}_s\mathbf{X}\mathbf{T}^{1/2}\left(\mathbf{T}^{1/2}\mathbf{X}'\left(\mathbf{\Delta}_s\mathbf{U}\mathbf{\Delta}_s\right)^{\dagger}\mathbf{X}\mathbf{T}^{1/2}\right)^{\dagger}$, then we will have $\mathbf{M}\mathbf{\Delta}_s\boldsymbol{\alpha} = \mathbf{M}\boldsymbol{\alpha} = \mathbf{0}$.

References

Ben-Israel, A., and Greville, T.N.E. (2002). *Generalized Inverses: Theory and Applications (Second Ed.)*. New York: Springer-Verlag.

Cassel, C.-M., Särndal, C.-E. and Wretman, J.H. (1977). *Foundations of Inference in Survey Sampling*. New York: John Wiley & Sons, Inc.

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

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

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.

Lunn, A.D., and Davies, S.J. (1998). A note on generating correlated binary variables. *Biometrika*, 85, 487-490.

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

Théberge, A. (1999). Extensions of calibration estimators in survey sampling. *Journal of the American Statistical Association*, 94, 635-644.

Théberge, A. (2017). Estimation when the covariance structure of the variable of interest is positive definite. *Journal of Official Statistics*, 33, 275-299.

Survey Methodology, June 2022
Vol. 48, No. 1, pp. 191-224
Statistics Canada, Catalogue No. 12-001-X

Is undesirable answer behaviour consistent across surveys? An investigation into respondent characteristics

Frank Bais, Barry Schouten and Vera Toepoel[1]

Abstract

In this study, we investigate to what extent the respondent characteristics age and educational level may be associated with undesirable answer behaviour (UAB) consistently across surveys. We use data from panel respondents who participated in ten general population surveys of CentERdata and Statistics Netherlands. A new method to visually present UAB and an inventive adaptation of a non-parametric effect size measure are used. The occurrence of UAB of respondents with specific characteristics is summarized in density distributions that we refer to as respondent profiles. An adaptation of the robust effect size Cliff's Delta is used to compare respondent profiles on the potentially consistent occurrence of UAB across surveys. Taking all surveys together, the degree of UAB varies by age and education. The results do *not* show consistent UAB across individual surveys: Age and educational level are associated with a relatively *higher* occurrence of UAB for some surveys, but a relatively *lower* occurrence for other surveys. We conclude that the occurrence of UAB across surveys may be more dependent on the survey and its items than on respondent's cognitive ability.

Key Words: Respondent profiles; Answer behaviour consistency; Adapted Cliff's Delta; Measurement error; Cognitive ability; Satisficing.

1. Introduction

The relation between answer behaviour in surveys and measurement error has been studied extensively. Measurement error refers to the extent to which a response deviates from the true value that a survey question was intended to measure (De Leeuw, Hox and Dillman, 2008). The occurrence and size of measurement error and hence response data quality can be influenced by respondent characteristics (Olson and Smyth, 2015; Tourangeau, Rips and Rasinski, 2000). Respondent characteristics can be thought of as fixed tendencies of a respondent that may lead to undesirable answer behaviour (UAB), like satisficing (Holbrook, Green and Krosnick, 2003; Kaminska, McCutcheon and Billiet, 2010). When respondents satisfice, they take short-cuts in the question-answering process. Satisficing can be seen as the outcome of the interaction of question difficulty, motivation, and cognitive ability (Krosnick, 1991, 1999; Krosnick, Narayan and Smith, 1996). Cognitive ability may be considered a characteristic of the respondent that is relatively constant over time. A straightforward proxy for cognitive ability like age or educational level may be used as a background variable to investigate its relation to answer behaviour. Background variables may not be free of measurement errors themselves, but these errors are assumed not to relate to answer behaviour and to be relatively stable over time (Schouten and Calinescu, 2013).

Answer behaviour should be stable and typical for the respondent in order to investigate its relation to respondent characteristics. That is, the behaviour for a specific respondent must be shown *consistently* in

1. Frank Bais, PO Box 1034, 6801 MG Arnhem. E-mail: frank.bais@cito.nl; Barry Schouten, PO Box 24500; 2490 HA Den Haag. E-mail: bstn@cbs.nl; Vera Toepoel, PO Box 80140, 3508 TC Utrecht. E-mail: v.toepoel@uu.nl.

order to be typical for that respondent. Here, the term "consistent" refers to a pattern of answer behaviour that is shown over several moments in time, across multiple surveys. When a respondent only incidentally shows a specific answer behaviour, it is not to say whether this is typical for that specific respondent. For instance, a respondent could fill out a single battery or set of five multiple choice items by choosing the very first answering option for each item. It is however not clear to what extent this may be a form of satisficing (Krosnick, 1991, 1999; Krosnick et al., 1996), as the answers may just as well be truly applicable to that respondent. In case of consistent answer behaviour, we may connect the behaviour to other stable characteristics of the same respondent. *In this paper, we investigate the relation between cognitive ability and consistent undesirable answer behaviour.* For this purpose, we use the respondent background variables age and educational level as proxies for cognitive ability. From here, we use the abbreviation "UAB" for the term "undesirable answer behaviour" throughout the paper.

Investigating the relation between cognitive ability and UAB is not new. However, this relation has not previously been investigated for a large sample of panel respondents across many surveys. To empower finding potential consistency for types of respondents in showing specific UAB, we use data from ten large population surveys administered by CentERdata in the LISS Panel. These surveys vary broadly in topic and contain many different kinds of items. By including many different surveys, variation will be present in survey topic and design. As a result of this variation, we assume that each survey has its own specific effect on the UABs. In our study, we want to distinguish respondent UAB that is survey-specific from UAB that occurs consistently across surveys. In order for respondent consistency to appear, UAB needs to occur across topics and survey designs. In other words, we need the full presence of topic and design variability to investigate UAB consistency across various surveys. We consider this topic and design variability as given and do not take into account survey and item characteristics for this study.

This study aims at linking cognitive ability to measurement error by using our method of constructing behaviour profiles. In case cognitive ability appears to have a consistent relation to specific UABs, surveys can be adapted according to the age or educational level of respondents in order to minimize measurement error. In case of such structural associations, the adaptation can be done globally, regardless of the survey. This also implies that our method could be used to predict measurement error. This means that time-consuming and expensive tests that examine the risk of measurement error could initially be omitted. If our method shows an increased risk of measurement error for specific respondents, setting up such tests could be valuable. If our method does not find such an increased risk, we could conclude that survey-independent adaptive survey design based on cognitive ability may not be useful.

For the purpose of our study, the specific survey topic or design would not even have to be taken into account. We realize that examining item characteristics and other respondent characteristics on their relation to measurement error across surveys is relevant as well. However, we consider our study a first step into investigating characteristics of respondents and items in their potentially consistent relation to UAB and measurement error across surveys. For this first step, we chose to examine the obvious respondent characteristics age and educational level in relation to eight relevant UABs (see Section 2).

Note that the undesirability of answer behaviour is potential by definition as we cannot validate its truthfulness (see Bais, Schouten and Toepoel, 2020 for an elaboration). Considering the aforementioned example, filling out the first answering option for all five items of a battery may refer to satisficing or to truthful responses. In the case of satisficing, we could say that this answer behaviour is undesirable. In the case of truthful responses, the behaviour is not undesirable. Our idea is that answer behaviour may refer to being undesirable as it is consistently shown across more surveys. The more consistent the behaviour, the more likely it becomes that the respondent is showing a personal pattern or style, and the more undesirable the behaviour may be considered. Therefore, the term "undesirable" is inherently potential when used throughout this paper. In summary, our foundation of ten large different surveys to detect potential behaviour consistency and to indicate the extent to which behaviour may be undesirable is solid and powerful.

This paper reads as follows: In Section 2 of this paper, we briefly elaborate on the theoretical framework on which our main research question is based. In Section 3, we describe the data, methods, and statistics that were used to compare the different age and educational categories for each UAB across surveys. As a method to detection of consistent UAB, we use so-called "respondent profiles", as suggested and explored by Bais (2021). In Section 4, we show all statistical results and give answers to our main research question. In Section 5, we conclude with a discussion of these results and make suggestions on how to proceed.

2. Theoretical framework

Cognitive ability may be considered a stable personal characteristic that has its influence on UAB (Krosnick, 1991, 1999; Krosnick et al., 1996). For our study, we consider the respondent characteristics age and educational level as proxies for cognitive ability to investigate its relation to specific UAB. Both age and educational level have been shown to be related to UAB and hence survey data quality (Krosnick, 1991, 1999; Krosnick et al., 1996). Older and lower educated respondents show less accurate UAB than younger respondents (Andrews and Herzog, 1986) and higher educated respondents (Antoni, Bela and Vicari, 2019), and a less stable attitude reliability measurement than younger and higher educated respondents (Alwin and Krosnick, 1991). See Table 2.1 for an overview of the age and educational categories as used in this study, and relevant literature.

In this study, we include two overarching kinds of UAB: Satisficing behaviour, and behaviour that is based on sensitive content. Satisficing behaviour refers to taking short-cuts in the question-answering process. Satisficing is positively related to item difficulty and can be the outcome of low cognitive ability (Heerwegh and Loosveldt, 2011; Krosnick, 1991, 1999; Krosnick et al., 1996). As a result of satisficing, respondents may show one of the following six specific UABs: Answering "don't know", acquiescence, neutral responding, extreme responding, primacy responding, and straightlining. See Table 2.2 for the meaning of these UABs and their relevant literature.

UAB can also be the result of sensitive survey content. Such UAB is positively related to item sensitivity and may be the outcome of a lack of willingness from the respondent to give a true answer (Bradburn, Sudman, Blair and Stocking, 1978; Shoemaker, Eichholz and Skewes, 2002; Tourangeau et al., 2000). Sensitive items may involve a threat of disclosure (Lensvelt-Mulders, 2008) or can be experienced as intrusive (Tourangeau et al., 2000; Tourangeau and Yan, 2007). As a result of sensitive content, respondents may give one of the following two specific UABs: Socially desirable responding and answering "won't tell". Note that "socially desirable responding" is in fact undesirable because of its relation to measurement error (see for instance DeMaio, 1984; Heerwegh and Loosveldt, 2011). See Table 2.2 for the meaning of the UABs and relevant literature. See Figure 2.1 for the complete theoretical framework.

We need to emphasize that most of the specific UABs in this study are referred to in some literature as "response styles" (see for instance He and Van de Vijver, 2013; He, Van de Vijver, Espinosa and Mui, 2014; Van Herk, Poortinga and Verhallen, 2004; Van Rosmalen, Van Herk and Groenen, 2010). We deliberately do not use the concept of response style throughout this paper. The goal of this study is to investigate whether groups of respondents express a stable and consistent pattern or style of specific UAB across surveys. This means that we need to avoid confusing "response style" as a UAB with "style" as a consistent pattern that groups may show across surveys. Therefore, we distinguished between the UAB itself and the pattern or style of UAB across surveys that we are actually expecting to find.

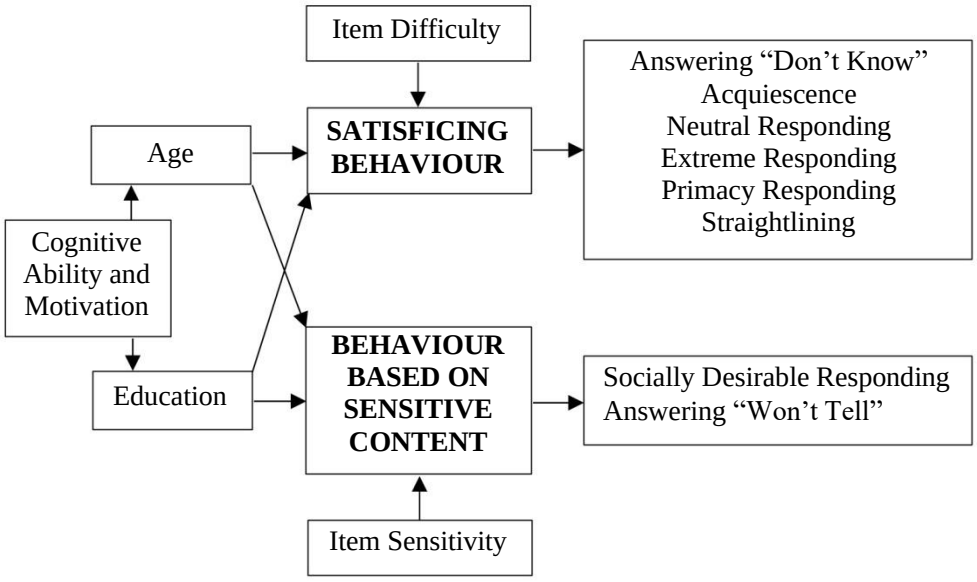
Table 2.1
Respondent characteristics, their categories, and selected relevant literature

Respondent characteristic	Categories of the respondent characteristics in this study	Selected relevant literature
Age	1. 15-24 years old 2. 25-34 years old 3. 35-44 years old 4. 45-54 years old 5. 55-64 years old 6. 65 years and older	Alwin and Krosnick (1991); Andrews and Herzog (1986); Greenleaf (1992); He, Van de Vijver, Espinosa and Mui (2014); Hox et al. (1991); Kieruj and Moors (2013); Meisenberg and Williams (2008); O'Muircheartaigh, Krosnick and Helic (2000); Pickery and Loosveldt (1998); Schonlau and Toepoel (2015); Zhang and Conrad (2014)
Education	1. primary school 2. vmbo: intermediate secondary education 3. havo/vwo: higher secondary education 4. mbo: intermediate vocational education 5. hbo: higher vocational education 6. wo: university	Aichholzer (2013); Alwin and Krosnick (1991); Greenleaf (1992); He et al. (2014); Krosnick (1991); Krosnick and Alwin (1987); Krosnick, Holbrook, Berent, Carson, Hanemann, Kopp, Mitchell, Presser, Ruud, Smith, Moody, Green and Conaway (2002); Marín, Gamba and Marín (1992); McClendon (1986, 1991); Narayan and Krosnick (1996); O'Muircheartaigh et al. (2000); Pickery and Loosveldt (1998); Schuman and Presser (1981); Zhang and Conrad (2014)

Table 2.2
The answer behaviours, their meaning, and selected relevant literature

Answer Behaviour	Meaning of the Answer Behaviour	Selected Relevant Literature for the Answer Behaviour
Socially Desirable Responding	The tendency to minimize showing socially undesirable behaviour.	Andersen and Mayerl, 2019; Campanelli, Nicolaas, Jäckle, Lynn, Hope, Blake and Gray, 2011; DeMaio, 1984; Heerwegh and Loosveldt, 2011; Holbrook et al., 2003; Jann, Krumpal and Wolter, 2019; Johnson and Van de Vijver, 2003; Kreuter, Presser and Tourangeau, 2008; Krosnick, 1999; Paulhus, 2002; Roberts, 2007; Roberts and Jäckle, 2012; Tourangeau et al., 2000; Tourangeau and Yan, 2007
Answering "Don't Know" and "Won't Tell"	The tendency to give a "don't know"- or a "won't tell"- answer to a question.	Beatty and Herrmann, 2002; Binswanger, Schunk and Toepoel, 2013; Bishop, Tuchfarber and Oldendick, 1986; Bradburn et al., 1978; Fricker, Galesic, Tourangeau and Yan, 2005; Krosnick et al., 2002; Leigh and Martin, 1987; Roberts, 2007; Roßmann, Gummer and Silber, 2017; Schuman and Presser, 1981; Shoemaker et al., 2002; Tourangeau et al., 2000; Vis-Visschers, Arends-Tóth, Giesen and Meertens, 2008
Acquiescence	The tendency to answer affirmatively, regardless of the content of the question.	Billiet and McClendon, 2000; De Leeuw, 1992; Díaz de Rada and Domínguez, 2015; Heerwegh and Loosveldt, 2011; McClendon, 1991; Messick, 1966; O'Muircheartaigh et al., 2000; Saris, Revilla, Krosnick and Shaeffer, 2010; Schaeffer and Presser, 2003; Stricker, 1963
Neutral Responding	The tendency to choose the neutral midpoint category from a bipolar answering scale.	He and Van de Vijver, 2013; Kalton, Roberts and Holt, 1980; Krosnick and Fabrigar, 1997; O'Muircheartaigh et al., 2000; Si and Cullen, 1998; Stern, Dillman and Smyth, 2007; Tarnai and Dillman, 1992
Extreme Responding	The tendency to choose an extreme category from the answering scale.	Aichholzer, 2013; De Leeuw, 1992; Díaz de Rada and Domínguez, 2015; Ye, Fulton and Tourangeau, 2011
Primacy Responding	The tendency to choose an option at the beginning of an answering list.	Galesic, Tourangeau, Couper and Conrad, 2008; Krosnick, 1991; Krosnick, 1992; Krosnick and Alwin, 1987; McClendon, 1991; Stern et al., 2007
Straightlining	The tendency to give the same answers to a series of questions arranged in a grid format.	Díaz de Rada and Domínguez, 2015; Fricker et al., 2005; Krosnick, 1991; Krosnick and Alwin, 1989; Roßmann et al., 2017; Schonlau and Toepoel, 2015; Zhang, 2013; Zhang and Conrad, 2014

Figure 2.1 Literature-based theoretical framework.



Literature overview: Age and education

Age and education seem to be related to non-substantive UAB, giving neutral, extreme, and acquiescent answers, and straightlining. Some studies found more acquiescence for older than for younger respondents (Meisenberg and Williams, 2008; O'Muircheartaigh, Krosnick and Helic, 2000), while other studies found the opposite (Hox, De Leeuw and Kreft, 1991) or no effect (He, Van de Vijver, Espinosa and Mui, 2014). Older respondents are found to give more extreme answers (Greenleaf, 1992; He et al., 2014; Meisenberg and Williams, 2008), including across questionnaires (Kieruj and Moors, 2013), while younger respondents are found to choose relatively more middle or neutral options (He et al., 2014). Schonlau and Toepoel (2015) found more straightlining for younger than for older respondents, while another study did not find a relation between age and straightlining for respondents who give answers at a high pace (Zhang and Conrad, 2013). Older respondents are found to give more "no opinion"-answers (Pickery and Loosveldt, 1998) or "don't know"-answers (O'Muircheartaigh et al., 2000) than younger respondents.

Lower educated respondents are found to give more "no opinion"-answers (Narayan and Krosnick, 1996; Krosnick et al., 2002; Pickery and Loosveldt, 1998) and "don't know"-answers (O'Muircheartaigh et al., 2000; Schuman and Presser, 1981) than higher educated respondents. Most studies found a negative relation between education and acquiescence (McClendon, 1991; Narayan and Krosnick, 1996; O'Muircheartaigh et al., 2000), although some research did not find a relation (Bachman and O'Malley, 1984; He et al., 2014; Hox et al., 1991). Also a negative relation between education and extreme responding is found (Aichholzer, 2013; Greenleaf, 1992; He et al., 2014; Marín, Gamba and Marín, 1992 – but see Bachman and O'Malley, 1984 for different findings), while mixed results exist concerning choosing middle or neutral options; see Narayan and Krosnick (1996) versus He et al. (2014). Among

respondents who give answers at a high pace, more straightlining was found for lower than for higher educated respondents (Zhang and Conrad, 2013). Evidence for the relation between education and primacy responding was mixed; see Krosnick and Alwin (1987) versus McClendon (1991).

As summarized above, the literature shows that the relation between age or education and UAB is not unambiguous. The literature needs to be complemented by results that are based on a fixed panel of respondents filling out multiple surveys. Existing findings from different studies are often mixed and may not be comparable because of different respondent samples. This means that it is hard to make literature-based predictions for our panel study and consistent UAB across surveys. Therefore, we do not construct hypotheses and merely explore to what degree UAB for different age and educational groups is consistent across surveys. By using a fixed panel and large set of ten surveys, our aim is to obtain an overarching overview of the relation of age and education to eight relevant UABs.

3. Method

3.1 LISS panel and surveys

We selected ten Dutch general population surveys that were administered by CentERdata to respondents of the Longitudinal Internet studies for the Social Sciences (LISS) Panel. This was done in the time period between June 2012 and December 2013. The surveys were the first wave of the Dutch Labour Force Survey from Statistics Netherlands and nine of the core studies from CentERdata. The data for the background variables as presented in Section 2 were also provided by CentERdata. All surveys were administered in computer-assisted format. The ten surveys cover a broad range of topics in the field of general population statistics, see Table 3.1. Also note the relatively high response rates for all surveys, ensuring comparable samples across the surveys. Considering these high and comparable response rates, we do not expect them to have a substantial relation to the occurrence of UAB within the context of this study.

The LISS Panel consists of about 7,000 individuals from about 4,500 households and is based on a probability sample of households. This sample is drawn from the population registry by Statistics Netherlands. All panel members were invited for all surveys included in this study. The first administration period for each survey was approximately a month. In case of initial nonresponse, the respondent was sent one or two reminders within this period. To increase the response rate, a second administration period of about a month including one or two reminders was executed for each survey. The respondents were compensated for each survey that they completed. This whole procedure was standardized for each survey, ensuring the comparability of the response rates for the surveys. The number of respondents that filled out a specific survey differed per survey and the number of surveys that respondents filled out varied across respondents. The average number of surveys filled out by a

respondent was almost eight. Altogether, the surveys contain 2,074 items that were used to cover the UABs as presented in Section 2.

Table 3.1
Overview of all surveys, a description of their content, and their response rate (and the number of respondents)

Survey (administration period, nr. of items)	Topics of the content	Response rate (and nr. of respondents)
Economic Situation Assets (AS) (Jun/Jul '12, i = 50)	Income, property and investment	75.2% (5,588)
Family and Household (FA) (Mar/Apr '13, i = 409)	Housing and household; social behaviour	88.8% (5,826)
Health (HE) (Nov/Dec '12, i = 243)	Health and well-being	85.4% (5,780)
Economic Situation Housing (HO) (Jun/Jul '13, i = 73)	Housing and household; income, property and investment	58.2% (3,199)
Economic Situation Income (IN) (Jun/Jul '13, i = 286)	Employment, labour, retirement; income, property, investment; social security, welfare	78.4% (5,015)
Personality (PE) (May/Jun '13, i = 200)	Psychology	90.6% (5,169)
Politics and Values (PO) (Dec '12/Jan '13, i = 148)	Politics; social attitudes and values	85.7% (5,732)
Religion and Ethnicity (RE) (Jan/Feb '13, i = 71)	Religion; social stratification and groupings	88.6% (5,908)
Work and Schooling (WO) (Apr/May '13, i = 471)	Education; employment, labour and retirement	86.5% (5,585)
Labour Force Survey (LF) (Dec '13, i = 123)	Education; employment and labour	81.2% (3,166)

3.2 Coding the undesirable answer behaviours

Each item (the total of the question and all answering options together) of all surveys was investigated on whether it was eligible for the selected UABs separately. The answering categories of the eligible items were coded for each UAB. In case a category was filled out for which the UAB occurred, the response was coded as 1; in case a category was filled out for which the UAB did not occur, the response was coded as 0. For all UABs, the coding was relatively straightforward. For neutral responding and answering "don't know" and "won't tell", the neutral, don't know- and won't tell-options respectively were coded as 1, while all other options were coded as 0. For extreme responding, the most negative and most positive option were coded as 1, while all other options were coded as 0. For primacy responding, the first two options were coded as 1, while all other options were coded as 0. This coding method was based on Medway and Tourangeau (2015) for the UABs that matched our research. See Table 3.2 for an overview of the UABs and their eligible kind of items. See Table 3.3 for the proportions of items for which the UABs are applicable per survey and in total. From here, we discuss the coding process of the UABs that need more elaboration: Socially desirable responding, acquiescence, and straightlining.

Table 3.2
The answer behaviours and their eligible kind of items

Answer Behaviour	Eligible items
Defined on Item Level	
Socially Desirable Responding	All items coded as asking for sensitive information, containing at least one answer category coded as possibly being socially desirable and at least one category coded as not being socially desirable.
Answering "Don't Know"	All items containing a "don't know" answer category.
Answering "Won't Tell"	All items containing a "won't tell" answer category.
Acquiescence	All more or less subjective (battery) items in the form of an ordinal agree/disagree or yes/no answer scale.
Neutral Responding	All (battery) items with an odd and minimum number of five answer categories on an ordinal scale, containing a neutral middle answer category.
Extreme Responding	All (battery) items with a minimum number of four answer categories on an ordinal scale, containing non-neutral first and last answer categories.
Primacy Responding	All (battery) items containing at least four response options.
Defined on Battery Level	
Straightlining	The items of all batteries containing at least 3 items and at least 4 answer categories, only in case all items of the battery were actually filled out.

Table 3.3
The number of items and batteries per survey, the average number of items per battery, and the proportions of items for which the answer behaviours are applicable for all surveys and in total*

	AS	FA	HE	HO	IN	PE	PO	RE	WO	LF	TO
Nr. of items	50	409	243	73	286	200	148	71	471	123	2,074
Nr. of batteries	-	11	5	-	3	16	12	4	2	-	53
Ave. nr. of items/battery	-	5.5	7.6	-	5.7	11.1	6.0	5.8	12.0	-	7.8
Soc. Des. responding	0.20	0.12	0.62	0.01	0.25	0.30	0.51	0.42	0.19	0.32	0.28
Answering "don't know"	0.52	0.08	0.01	0.33	0.47	0.02	0.45	0.49	0.11	0.01	0.18
Answering "won't tell"	0.28	-	-	0.30	0.31	-	0.01	-	0.04	0.81	0.12
Acquiescence	-	0.03	-	-	0.01	0.96	0.68	0.24	0.05	0.03	0.17
Neutral responding	-	0.10	-	-	0.05	0.93	0.66	-	0.04	-	0.17
Extreme responding	-	0.13	-	-	0.05	0.93	0.66	-	0.06	-	0.18
Primacy responding	-	0.37	0.23	-	0.24	0.93	0.73	0.55	0.19	0.27	0.35
Straightlining	-	0.15	0.16	-	0.06	0.89	0.49	0.32	0.05	-	0.20

*Assets (AS), Family (FA), Health (HE), Housing (HO), Income (IN), Personality (PE), Politics (PO), Religion (RE), Work (WO), Labour Force Survey (LF), Total (TO).

Socially desirable responding

About 50% of all items of the involved surveys together were coded as potentially asking for sensitive information by at least one of three coders (see Bais, Schouten, Lugtig, Toepoel, Arends-Tóth, Douhou, Kieruj, Morren and Vis, 2019). Next, the answering categories of these items were coded by an independent fourth coder on whether they may refer to a socially desirable answer. Let us consider the following example:

"Can you indicate, on a scale from 0 to 10, how hard or how easy it is for you to live off your income?

0 means that it is very hard to live off your income, 10 means that it is very easy.

very hard very easy

0 1 2 3 4 5 6 7 8 9 10"

The idea is that it is socially desirable to state that it is relatively easy to live off one's income. For our study, we only considered the answering options 8 through 10 as socially desirable options. In this way, we hoped to distinguish respondents who are clearly sensitive to responding in a socially desirable manner across surveys from those who are not.

Acquiescence: Responding agreeably/affirmatively to a question

The answering categories of all items were evaluated on whether they showed an extent of agreeableness or affirmativeness (see Medway and Tourangeau, 2015). Both positively and negatively worded items were present throughout the surveys to measure acquiescence. Both battery (a set of related items sharing the same answering options) and non-battery items were considered and also subjective variants of the typical answering option "agree", like "satisfied", "applicable", and "yes", were considered for acquiescence. We chose to include those variants as acquiescent options to capture a broad range of possible acquiescent behaviour across many items. Such a broad range may result into more variation between respondents in showing acquiescence, so that we may better distinguish acquiescent from non-acquiescent respondents. Let us consider the following example:

"I really enjoy responding to questionnaires through the mail or Internet.

totally disagree totally agree

1 2 3 4 5 6 7"

For our study, we considered the answering options 5 through 7 as acquiescent options. We decided to consider the option "somewhat agree" (option 5 in the example) as an acquiescent response as well, as we hoped to distinguish respondents who acquiesce clearly or to only a certain extent from respondents who do not acquiesce.

We need to note that the coding of socially desirable responding and acquiescence is more or less arbitrary; the coding of both UABs may have been executed either more or less strictly. On the one hand, this means that a response option that was coded as socially desirable or acquiescent may be a socially desirable or acquiescent response for some respondents, but the intended response for others. On the other hand, a response option that was *not* coded as socially desirable or acquiescent may indeed be the intended response for some respondents, but should have been coded as socially desirable or acquiescent for others. However, in order to investigate socially desirable responding and acquiescence at all, a coding threshold that distinguishes the occurrence from the non-occurrence of these UABs simply needs to be placed at some point. By the current way of coding these UABs, enough variability between respondents is present in order to distinguish age and educational subgroups that may differ in the occurrence of UAB.

Straightlining: Choosing the same answering category for all items in a battery

Our idea is to consider straightlining for a battery only when the very same answering options were filled out *for all its items* (see Schonlau and Toepoel, 2015). When this is the case, the number of times that a "1" is coded is equal to the number of items that the battery consists of. For instance, the occurrence of straightlining for a battery of five items received the code "1" five times. This means that we took into account the length of the battery for this UAB. In other words, the more items a battery consists of, the stronger the UAB refers to straightlining in case a respondent filled out the same option for each item. See the following section for an elaboration on how the coding at the item level for all UABs is transformed into meaningful respondent behaviour summaries.

3.3 Respondent profiles

In order to compare respondents on consistent UAB across surveys, a few aspects need to be taken into account regarding the UAB. First, the number of items that is applicable to the UAB per survey can be relatively small. This means that uncertainty exists around the actual occurrence of UAB, since it is based on, by definition, a limited number of items per respondent. To give an example, suppose a respondent A fills out ten items and gives a "don't know"-answer five times, while another respondent B fills out 100 items and gives a "don't know"-answer 50 times. Although both respondents can be attributed a probability of 0.50 for answering "don't know", this probability is relatively more certain for respondent B since it is based on more response data. In other words, the actual occurrence of UAB for respondents may be more uncertain as respondents fill out a smaller number of items.

Second, when a survey contains filter questions that may or may not branch out into follow-up questions, each respondent is likely to fill out a different number of items for that survey. Therefore, the actual occurrence of UAB is indicated with varying uncertainty across different respondents within a survey. Hence, to compare respondents sharing the same characteristic on their UAB across surveys, simply using individual UAB proportions is insufficient: A method must be used that takes into account these uncertainties. For this purpose, we introduce the method of using respondent profiles. See Bais (2021) for an extensive statistical elaboration on this method.

The respondent profile

The respondent profile is a summary of UAB for a group of respondents. It represents the relative proportions of a specified population group (for instance lower educated respondents) in showing a specified UAB (for instance answering "don't know") at all possible probabilities from 0 to 1. In constructing a respondent profile, we make use of the binomial distribution to take into account the abovementioned uncertainties. Note that when we speak of a "respondent profile", we refer to a group of respondents by definition. When we discuss a profile for a single respondent, we explicitly speak of an "individual respondent profile".

Consider an individual respondent r who fills out a survey consisting of 50 items of which each offers the answering option "don't know". Suppose that the respondent chooses the "don't know"-option 10 times out of the 50 possible occasions. Then these numbers are used to construct a binomial distribution. This binomial distribution shows the occurrence of answering "don't know" for respondent r. The likelihood of the UAB occurrence is calculated for each probability along the probability range from 0 to 1. For practical calculation, we chose for a probability step size interval of 0.01 in order to construct the binomial distribution on the basis of 100 probabilities. We call the resulting binomial distribution for respondent r an individual respondent profile. An individual respondent profile is the likelihood curve for the UAB occurrence and is calculated for each probability from 0 to 1. Hence, to construct the individual profile for respondent r, the likelihood of the UAB occurrence is calculated on the basis of 10 actual "don't know"-answers out of 50 possible occasions for all 100 probabilities:

$$\lambda_r(p) = \binom{I_r}{G_r} p^{G_r} (1-p)^{I_r - G_r}, \tag{3.1}$$

where λ_r is the likelihood curve or individual profile for respondent r, p is the probability between 0 and 1 with step size 0.01, I_r is the number of items for which choosing the UAB is possible for respondent r, and G_r is the number of items for which the behaviour is actually shown by respondent r. In order to make individual respondent profiles comparable, we normalize the resulting distribution to obtain an area below the curve of 1 regardless of step size. This is done by dividing each of the likelihoods that the profile consists of by the sum of all likelihoods:

$$\tilde{\lambda}_r(p) = \frac{\lambda_r(p)}{\int_{p=0}^{1} \lambda_r(p)\,dp}, \tag{3.2}$$

where $\tilde{\lambda}_r$ is the normalized individual profile for respondent r. For a single respondent r, the average or expected value E_r for the UAB occurrence can be estimated on the basis of the respondent's profile and the integral over p. This means that each probability from 0 to 1 is multiplied by its accompanying likelihood:

$$E_r = \int_{p=0}^{1} p\tilde{\lambda}_r(p)\,dp. \tag{3.3}$$

The likelihood curve resulting from formula's (3.1) and (3.2) is an individual respondent profile. The profile delineates the expected UAB occurrence across the full potential probability range from 0 to 1 and gives consideration to the amount of occurrence uncertainty. To illustrate the uncertainty on the individual level, consider two respondents who may both have an expected UAB value of 0.50, but who filled out a different number of items for which the UAB was possible. For instance, respondents A and B showed UAB for 10 out of 20 items and for 30 out of 60 items respectively. See Graph 1 in Figure 3.1. Here, our method takes into account that the expected value of 0.50 is more precisely estimated for respondent B

than for respondent A. This is visible by the relatively more narrow and peaked profile for respondent B, indicating that this respondent's UAB occurrence is relatively more certain.

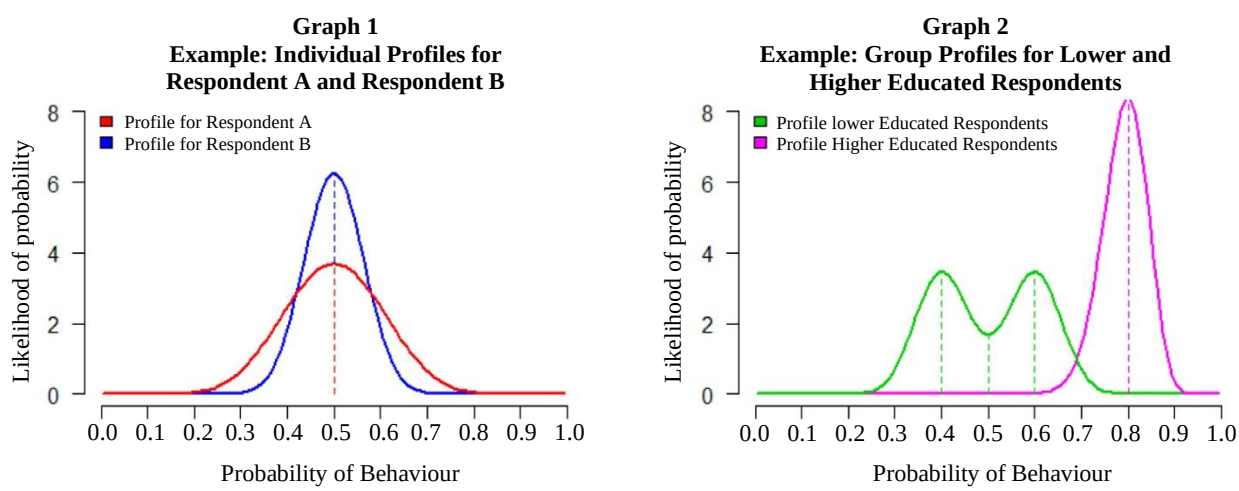
By considering all respondents who meet the condition of a specific category for a characteristic (for instance lower educated respondents for educational level), the average respondent group profile can be calculated by simply summing their comparable individual profiles and dividing the outcome by the number of respondents:

$$\bar{\lambda}(p) = \frac{1}{R}\sum_{r=1}^{R}\tilde{\lambda}_r(p), \tag{3.4}$$

where $\bar{\lambda}$ is the respondent profile of the group UAB occurrence averaged over all respondents, and R is the total number of respondents in the group. By means of this average respondent profile, the averaged expected value $\bar{E}$ for the UAB occurrence for this group of respondents can be calculated as follows:

$$\bar{E} = \int_{p=0}^{1} p\bar{\lambda}(p)dp. \tag{3.5}$$

Figure 3.1 Examples of respondent profiles with similar expected values (Graph 1) and different expected values (Graph 2).

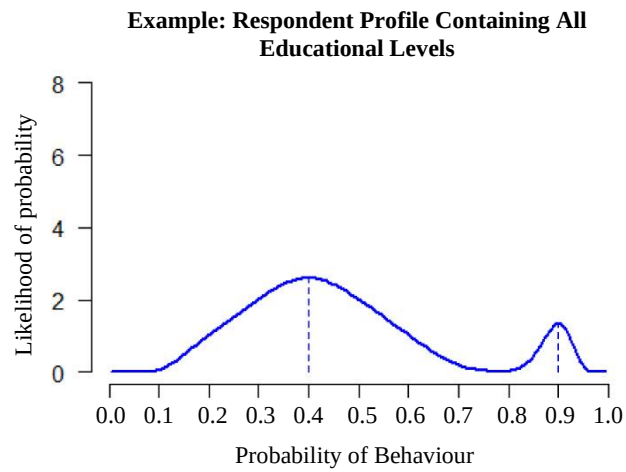


The likelihood curve resulting from formula (3.4) is a group respondent profile. To illustrate the uncertainty on the group level, consider the two groups of lower and higher educated respondents showing a specific UAB. See Graph 2 in Figure 3.1. The expected values for the groups are 0.50 and almost 0.80 respectively. Our method shows that the expected UAB occurrence is more precisely estimated for higher than for lower educated respondents. It is also visible that for lower educated respondents, the UAB occurrence is not centered around the expected group value of 0.50, but around the values of 0.40 and 0.60. Although formula (3.4) refers to a profile for a group of respondents, it does give an indication of

individual UAB. Consider the respondent profile in Figure 3.2 containing individuals on all educational levels. The majority of individuals does not show a specific UAB very often considering the large bump left of the center. On the right, a small peak is visible that refers to a subgroup of individuals showing the UAB very often. These respondents may be either lower or higher educated respondents, or they may share another characteristic that is associated with a high UAB occurrence. The point here is that the respondent profile takes into account the individual UAB and that subgroups of individuals showing a specific occurrence of UAB may be identified in the profile.

Note that by using this method of constructing respondent profiles, we assume that individual UAB is independent across items. This assumption may be partly unjustified, as there may be interdependence across items to some extent in practice. Elaborating on taking into account interdependence across items is beyond the scope of this paper. We refer to Bais (2021) for suggestions on how to cope with interdependence across items in future research using respondent profiles.

Figure 3.2 Example of a respondent profile containing all educational levels.



Also note that we choose not to use a more traditional model like multilevel analysis to analyze our data. We do not follow identified individual respondents across surveys, but we analyze subgroups of respondents sharing the same characteristic by our profile method for several reasons. Besides taking into account the uncertainty that comes along with the delimited and varying number of respondents and/or items, respondent profiles fully summarize and graphically visualize UAB for subgroups of respondents. And by means of full respondent profiles, relatively small subgroups that deviate from the main body of a larger group may be detected. Throughout this paper, note that a *category* of respondents refers to respondents in a specific single age or educational category (see Table 2.1), while a *(sub)group* of respondents may also refer to respondents from several age or educational categories.

In summary, the expected values of two groups with different characteristics indicate the average UAB occurrences for the groups as a whole. In this way, an idea is obtained about the difference of the occurrences of specific UAB (for instance answering don't know) between two groups (for instance lower and higher educated respondents). The next step is to use a solid analysis to compare the UAB occurrences of two groups.

3.4 Cliff's Delta for comparing groups of respondents

To compare two groups or categories of respondents meeting a specific characteristic, an adaptation of the effect size Cliff's Delta (Cliff, 1993, 1996ab) is used. Cliff's Delta δ can be used as a robust alternative to using two independent group means. Using Cliff's Delta for our research asks for an adapted version of the statistic, as we are not considering data observations but density distributions.

The original Cliff's Delta for data observations

Cliff's Delta δ is a robust effect size that indicates to what extent two groups are different. It calculates the probability that a random data observation X_a from a group A is larger than a random data observation X_b from another group B, minus the reverse probability (Hess and Kromrey, 2004; Rousselet, Foxe and Bolam, 2016; Rousselet, Pernet and Wilcox, 2017). In practice, this means that each data observation in group A is compared to each data observation in group B. Then a value is assigned to each such comparison. If an observation from group A is larger than an observation in group B, this value is 1. If an observation in group A is smaller than an observation in group B, this value is -1. If the observations in group A and B are equal, this value is 0. Then the total sum of all these values is divided by the total number of comparisons, giving Cliff's Delta. The smaller the overlap between the distributions of two groups, the more difference between the two groups. A Cliff's Delta of -1 or 1 indicates absence of overlap between two groups and a Cliff's Delta of 0 refers to group equivalence (Hess and Kromrey, 2004). The sample estimate of Cliff's Delta $\hat{\delta}$ is

$$\hat{\delta} = \frac{\sum_{a=1}^{A} \sum_{b=1}^{B} \operatorname{sgn}\left(X_a - X_b\right)}{AB}, \tag{3.6}$$

where $\left(X_a - X_b\right)$ results in a positive or negative number or 0, the sign function "sgn" transforms each positive number into 1 and each negative number into -1, and preserves each 0, and A and B are the sizes of group A and group B respectively.

Adapting Cliff's Delta for density distributions

We need to adapt the original Cliff's Delta for our respondent profiles that consist of likelihood distributions. Consider Cliff's Delta for which each specific observation from sample A is compared to each specific observation from sample B exactly once. This means that when an observation with a specific value from sample A occurs three times, this observation value is compared to all observations

from sample B three times as well. Therefore, we may regard both observations for each such comparison on its own as having a "frequency" or "weight" of 1. When we transpose this idea to respondent profiles, we may consider the UAB probabilities from 0 to 1 (with a specific step size interval) our "observations" and the likelihoods for each probability their "frequencies" or "weights".

$$\hat{\delta} = \frac{\sum_{a=1}^{A} \sum_{b=1}^{B} \operatorname{sgn}\left(P_a - P_b\right) \bar{\lambda}\left(P_a\right) \bar{\lambda}\left(P_b\right)}{\sum_{a=1}^{A} \sum_{b=1}^{B} \bar{\lambda}\left(P_a\right) \bar{\lambda}\left(P_b\right)}, \tag{3.7}$$

where P_a and P_b are the probabilities from 0 to 1 from group A and group B respectively, $\bar{\lambda}(P_a)$ and $\bar{\lambda}(P_b)$ are the averaged likelihoods of the probabilities P_a and P_b respectively, and A and B are the same number of step size intervals for both groups.

As a brief illustration, we calculate the adapted Cliff's Delta by means of formula (3.7) for the respondent profiles in Figure 3.1. Consider Graph 1. When comparing the profiles for respondent A to respondent B, Cliff's Delta is 0. Although the two profiles slightly differ, their shapes are symmetrically formed around the shared expected value of 0.50. This means that the various values in the denominator of formula (3.7) cancel each other out. Consider Graph 2. When comparing the profiles for lower to higher educated respondents, Cliff's Delta is -0.99. The profiles hardly overlap and the higher educated respondents clearly show more of some UAB than the lower educated respondents. The reason that Cliff's Delta is not exactly 1 can be explained by the very small part of overlap around the probability of 0.70 (see Graph 2). Note that the sign would change and Cliff's Delta would be 0.99 when we would compare higher to lower (instead of lower to higher) educated respondents.

For our study, we use the adaptation of Cliff's Delta in order to compare respondent profiles. The respondent profiles and this adaptation take into account the fact that each respondent fills out a delimited and different number of items (see Section 3.3). Cliff's Delta has many advantages with respect to answering our research question. Cliff's Delta makes no assumption about the shape of the underlying distribution (Cliff, 1993, 1996ab; Goedhart, 2016; Vargha and Delaney, 2000) and is robust in case of outliers or skewed or otherwise non-normal distributions (Goedhart, 2016). Cliff's Delta is easy to calculate, straightforward to interpret, and standardized, meaning different effect size categories can be distinguished (Goedhart, 2016; see Section 4.2 for these categories). For our adapted Cliff's Delta, relatively small or unequal sample sizes are no issue.

3.5 Confidence intervals for Cliff's Delta and statistics

For each Cliff's Delta, we use confidence intervals to refer to its amount of uncertainty. For a respondent characteristic, each Cliff's Delta is based on the comparison between the profile of a category and the overall profile of the remaining categories taken together. For a confidence interval, we bootstrap 10,000 category profiles and 10,000 overall profiles. We use the so-called empirical bootstrap method, as we cannot make assumptions about the profiles that are non-parametric by definition (see for instance

Dekking, Kraaikamp, Lopuhaä and Meester, 2005 for more on this bootstrap method). For each profile, respondents are randomly sampled with replacement and their individual profiles are averaged by means of formula (3.4). The number of sampled respondents is equal to the number of respondents in the category or overall group respectively. By means of these averaged bootstrap profiles, we calculate 10,000 Cliff's Delta's and rank them from low to high. Because of the large number of Cliff's Delta's in our study, we choose to use 99% confidence intervals. This means that we use the 51st and the 9,950th Cliff's Delta in the ranking to construct each confidence interval. In the results section, we show Cliff's Delta outcomes for the respondent characteristics and their categories for all UABs. Each Cliff's Delta is accompanied by its 99% confidence interval.

4. Results

In this section, we first show the Cliff's Delta's for all surveys together as if they were one large survey. Second, we consider the Cliff's Delta's per survey to give an indication about UAB consistency across surveys to answer our research question. All Cliff's Delta's are obtained by comparing each category profile to the combined profile of the remaining categories. For instance, this means that the profile for respondents aged 15-24 are compared to the profile for the respondents from all other age categories. We chose for this type of comparison, as we are interested in whether a specific subgroup deviates from the complete sample of respondents, considered representative regarding age and education, minus that subgroup.

First, we need to note that respondents varied in the number of surveys they filled out. Some respondents filled out only one or two surveys, while others filled out all or almost all surveys. Behaviour data for *every* survey that the respondent filled out were used for the analyses. For instance, if a respondent filled out the surveys Health, Income, and Personality, this respondent is included in the data analyses for all these surveys. Second, respondents are classified in one category for both age and education. This means that a respondent can be older than 64 years and highly educated, and is included in the data analyses for both characteristics. Hence, respondents are included in each survey and characteristic analysis that is applicable to them. From this, it should be clear that we do not analyze individual respondents in this study, but that we focus on *groups* of respondents sharing the same characteristic. The reason is that we want to relate UAB to characteristics that are known from the literature to affect UAB, rather than to isolate individuals and explore potentially related characteristics.

We consider an individual respondent profile based on less than five items non-informative and too imprecise to take into account. Therefore, for each respondent group profile, we only include respondents who filled out at least five items. This means that part of the respondents may be excluded from several subgroups for the analyses. As a result, the occurrence of UAB for a subgroup after excluding respondents may differ from the initial occurrence of UAB for that subgroup. Thus, after excluding respondents from a subgroup, the remainder of the subgroup may not be representative for the original subgroup anymore in

terms of the initial UAB occurrence. Therefore, we used two criteria to guarantee the representativeness of each original subgroup: 1) Each subgroup consists of more than 30% of the number of respondents in the original group, and; 2) the UAB occurrence in each subgroup does not differ more than 0.02 from the original group's UAB occurrence.

4.1 Exploring survey participation and respondents aged 65 or older

Before elaborating on the main results, we give the outcomes of a few explorations. First, we investigated to what extent frequency of survey participation may have differed between the various age and educational subgroups. See Table 4.1. The average number of surveys that was filled out per respondent overall is 7.6. The average number of surveys per educational subgroup appeared to be relatively high and not to differ much between subgroups. For the age subgroups however, it is evident that younger respondents filled out a lower number and older respondents a higher number of surveys on average.

Table 4.1
Overall survey participation in total and per subgroup in average number of surveys (and absolute number of respondents)

	TOT	15_24	25_34	35_44	45_54	55_64	> 64
Age	7.6 (6,700)	6.0 (838)	6.8 (803)	7.3 (1,083)	7.7 (1,223)	8.3 (1,289)	8.5 (1,464)
		Primary	VMBO	HAVW	MBO	HBO	WO
Education	7.6 (6,688)	7.3 (601)	7.7 (1,634)	7.3 (791)	7.6 (1,549)	7.7 (1,504)	7.6 (609)

We used respondent profiles and Cliff's Delta to explore whether the degree of participation made a difference in the occurrence of the specific UABs taking all surveys together. We split up the complete sample of panel respondents into a group who filled out at most eight surveys and a group who filled out at least nine surveys. See Table 4.2. It is clear that participation rate did not affect the occurrence of most UABs. Not surprisingly, respondents who participated in relatively few surveys showed relatively more "won't tell"-answers. A second effect was relatively more straightlining in case of a lower participation rate.

Table 4.2
Cliff's Delta for Low (Filled out at most eight surveys) versus High (Filled out at least nine surveys) survey participation per answer behaviour[1]

	SD	PR	DK	ST	WT	AC	NE	EX
At most eight vs. at least nine surveys	-0.09	0.07	0.08	*0.14* ~	*0.29* *	-0.06	0.02	-0.10

~→ small effect; *→ medium effect; #→ large effect.
[1]Socially Desirable Responding (SD), Primacy Responding (PR), Answering "Don't Know" (DK), Straightlining (ST), Answering "Won't Tell" (WT), Acquiescence (AC), Neutral Responding (NE), Extreme Responding (EX).

Lastly, respondents aged 75 or older may be even more vulnerable to difficulty in cognitive processing and hence showing UAB than respondents aged 65-74. Therefore, we compared respondents aged 65-74 to respondents aged 75 or older on their group UAB proportion. See Table A.1 in Appendix A. Age subgroups did not or hardly differ for most UABs and surveys. Only regarding straightlining there were a few striking differences, but interestingly, these showed that respondents aged 75 or older expressed *less* straightlining than respondents aged 65-74. This means that we do not have a reason to split up the age subgroup of 65 years or older into two smaller subgroups.

4.2 Overall outcomes for Cliff's Delta

The overall results for Cliff's Delta concern the global picture for specific subgroups for all surveys taken together. We use the rules that $|\delta| < 0.11$ indicates no effect, $0.11 \leq |\delta| < 0.28$ a small effect, $0.28 \leq |\delta| < 0.43$ a medium effect, and $|\delta| \geq 0.43$ a large effect, as investigated by Vargha and Delaney (2000), see also Goedhart (2016). A subgroup is always compared to the aggregated total of all remaining applicable subgroups regarding the specific characteristic. See Table 4.3 for the Cliff's Delta's for all surveys taken together.

From Table 4.3, it is clear that subgroups for age and education differ in various forms of specific satisficing behaviours overall. Younger and lower educated respondents showed more "don't know"-answers than older and higher educated respondents. Higher educated respondents showed more acquiescent, but less neutral responses than lower educated respondents. Younger respondents showed less extreme responses than respondents from other age categories. Respondents from the middle age categories showed more primacy responses than both younger and older respondents (see Graph 1 in Figure 4.1), while higher educated respondents showed more primacy responses than lower educated respondents. Respondents from the middle age categories showed more straightlining than older respondents, while higher educated respondents showed more straightlining than lower educated respondents. From Table 4.3, it is also evident that some subgroups for age and education differ for sensitivity-based answer behaviour overall. Younger respondents showed more "won't tell"-answers than older respondents. Higher educated respondents showed more socially desirable responses (see Graph 2 in Figure 4.1), but less "won't tell"-answers than lower educated respondents. In summary, overall satisficing and sensitivity-based behaviours are clearly present, in most cases particularly for the youngest, oldest, lowest educated, or highest educated respondent groups.

A present overall effect size for a specific category and UAB does not by definition mean a present effect size for various surveys; an overall effect size may exist without effect sizes for any surveys. The opposite may be true as well; an overall effect size may be absent, as positive and negative effect sizes for various surveys cancel each other out. In the following section, we investigate to what extent either positive or negative effect sizes consistently exist across surveys and answer our main research question.

Table 4.3
Overall Cliff's Delta (and its 99% confidence interval) taken over all surveys, for all age categories[1] and all educational categories[2] for all answer behaviours[3]

	Satisficing Behaviour						Behaviour Based on Sensitive Content	
	DK	**AC**	**NE**	**EX**	**PR**	**ST**	**SD**	**WT**
Age 1524	*0.30 ** (0.25, 0.35)	-0.06 (-0.12, -0.00)	-0.02 (-0.08, 0.04)	*-0.15 ~* (-0.21, -0.10)	*-0.24 ~* (-0.30, -0.18)	0.00 (-0.06, 0.07)	-0.04 (-0.09, 0.01)	*0.25 ~* (0.20, 0.31)
Age 2534	*0.11 ~* (0.05, 0.16)	0.05 (-0.00, 0.11)	-0.06 (-0.12, -0.00)	-0.08 (-0.14, -0.02)	0.08 (0.03, 0.14)	*0.12 ~* (0.06, 0.17)	0.02 (-0.03, 0.08)	0.09 (0.04, 0.14)
Age 3544	0.08 (0.04, 0.13)	-0.01 (-0.06, 0.04)	0.03 (-0.02, 0.07)	0.01 (-0.04, 0.06)	*0.13 ~* (0.08, 0.17)	*0.19 ~* (0.15, 0.24)	-0.02 (-0.07, 0.02)	0.08 (0.03, 0.12)
Age 4554	0.02 (-0.02, 0.07)	-0.04 (-0.09, 0.00)	0.01 (-0.04, 0.05)	0.04 (-0.01, 0.08)	*0.13 ~* (0.08, 0.17)	*0.11 ~* (0.07, 0.16)	-0.01 (-0.05, 0.03)	0.02 (-0.02, 0.06)
Age 5564	*-0.15 ~* (-0.19, -0.11)	0.03 (-0.01, 0.07)	-0.02 (-0.06, 0.02)	0.06 (0.01, 0.10)	0.06 (0.03, 0.10)	*-0.12 ~* (-0.16, -0.08)	0.02 (-0.02, 0.06)	-0.06 (-0.10, -0.02)
Age 65Ol	*-0.20 ~* (-0.24, -0.16)	0.02 (-0.02, 0.06)	0.04 (0.00, 0.08)	0.05 (0.01, 0.09)	*-0.17 ~* (-0.20, -0.13)	*-0.22 ~* (-0.26, -0.18)	0.02 (-0.02, 0.06)	*-0.17 ~* (-0.20, -0.14)
Edu PRI	*0.20 ~* (0.14, 0.26)	*-0.13 ~* (-0.19, -0.06)	*0.14 ~* (0.08, 0.20)	0.03 (-0.04, 0.10)	*-0.21 ~* (-0.27, -0.15)	*-0.14 ~* (-0.20, -0.07)	*-0.13 ~* (-0.20, -0.08)	0.08 (0.02, 0.14)
Edu VM	0.10 (0.06, 0.14)	*-0.18 ~* (-0.22, -0.14)	*0.14 ~* (0.10, 0.18)	0.04 (-0.00, 0.08)	*-0.13 ~* (-0.17, -0.09)	-0.04 (-0.08, 0.00)	-0.08 (-0.12, -0.04)	0.07 (0.04, 0.11)
Edu HA	0.00 (-0.05, 0.06)	0.01 (-0.04, 0.06)	-0.10 (-0.16, -0.05)	-0.02 (-0.08, 0.03)	0.00 (-0.05, 0.06)	-0.04 (-0.09, 0.02)	-0.06 (-0.10, -0.01)	0.02 (-0.03, 0.07)
Edu MB	0.07 (0.03, 0.11)	-0.04 (-0.08, 0.00)	0.05 (0.01, 0.09)	-0.02 (-0.07, 0.02)	0.02 (-0.02, 0.06)	0.05 (0.00, 0.09)	-0.02 (-0.06, 0.02)	0.08 (0.04, 0.11)
Edu HB	*-0.17 ~* (-0.21, -0.13)	*0.18 ~* (0.14, 0.22)	*-0.12 ~* (-0.16, -0.08)	-0.03 (-0.07, 0.01)	*0.12 ~* (0.09, 0.16)	0.02 (-0.02, 0.06)	*0.13 ~* (0.10, 0.17)	*-0.13 ~* (-0.16, -0.09)
Edu WO	*-0.21 ~* (-0.27, -0.16)	*0.22 ~* (0.16, 0.27)	*-0.18 ~* (-0.23, -0.12)	0.02 (-0.04, 0.08)	*0.19 ~* (0.14, 0.24)	*0.12 ~* (0.07, 0.18)	*0.14 ~* (0.08, 0.19)	*-0.13 ~* (-0.18, -0.08)

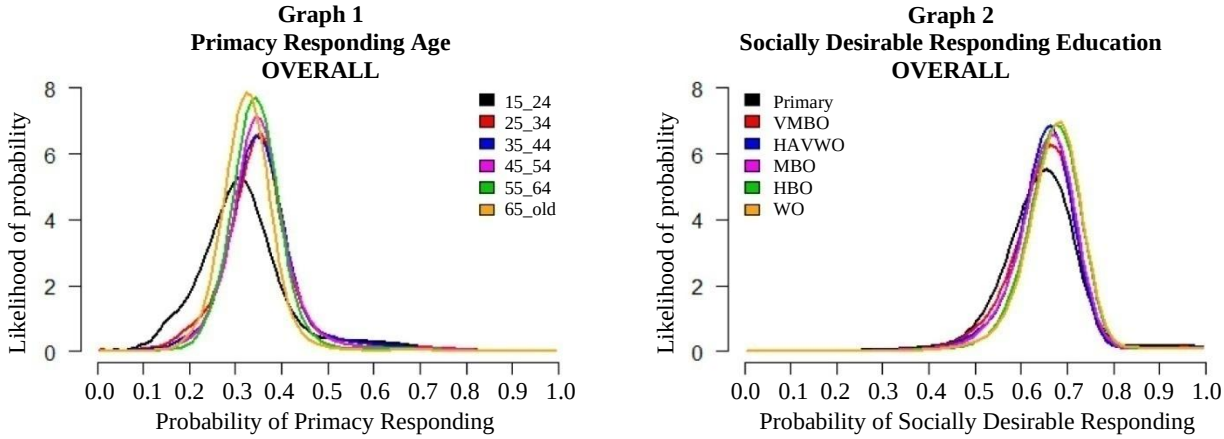
~→ small effect; *→ medium effect; #→ large effect.

[1]15-24 Years (Age 1524), 25-34 Years (Age 2534), 35-44 Years (Age 3544), 45-54 Years (Age 4554), 55-64 Years (Age 5564), 65 Years and Older (Age 65Ol).

[2]Primary Education (Edu PRI), VMBO (Edu VM), HAVWO (Edu HA), MBO (Edu MB), HBO (Edu HB), WO (Edu WO).

[3]Answering "Don't Know" (DK), Acquiescence (AC), Neutral Responding (NE), Extreme Responding (EX), Primacy Responding (PR), Straightlining (ST), Socially Desirable Responding (SD), Answering "Won't Tell" (WT).

Figure 4.1 Less Primacy Responding for Respondents Aged 15-24 (black) and 65 or Older (orange), and More Primacy Responding for Respondents Aged 35-44 (blue) and 45-54 (purple) in Graph 1; Less Socially Desirable Responding for Respondents Who Finished Only Primary School (black), and More Socially Desirable Responding for Respondents Who Finished HBO (green) or WO (orange) in Graph 2.



4.3 Consistency outcomes for Cliff's Delta

These results for Cliff's Delta concern the consistency of subgroups across surveys. To reveal consistency, we considered the number of surveys for which at least a small effect $\left(|\delta| \geq 0.11\right)$ was the result. Considering consistency conservatively, as an at least small effect for a specific UAB and category for *all or almost all* applicable surveys, we would draw the conclusion that there is no consistency to be found: *There is no consistent satisficing or sensitivity-based behaviour evident across surveys*. See Table 4.4 containing all results for the UABs and categories for which more than half of the applicable surveys showed either positive or negative effect sizes: There is no category that shows an effect for all or almost all surveys for any UAB.

Table 4.4
Cliff's Delta (and its 99% confidence interval) for the behaviours Answering don't know, Primacy responding, and Neutral responding, for the applicable Age categories[1] and Educational categories[2] for the Applicable surveys[3]

	FA	HE	HO	IN	PE	PO	RE	WO
Answering "Don't Know"								
Age 1524	0.09 (0.05, 0.12)			*0.46 #* (0.41, 0.51)		*0.28 ** (0.22, 0.34)	0.05 (0.03, 0.07)	*0.24 ~* (0.19, 0.30)
Age 65Ol			*-0.13 ~* (-0.17, -0.09)	*-0.20 ~* (-0.24, -0.16)		*-0.14 ~* (-0.17, -0.10)	-0.02 (-0.03, -0.01)	
Edu PRI	*0.15 ~* (0.08, 0.23)		0.08 (-0.00, 0.15)	*0.16 ~* (0.10, 0.23)		*0.17 ~* (0.11, 0.24)	0.02 (0.00, 0.05)	*0.23 ~* (0.15, 0.31)
Primacy Responding								
Age 1524	*-0.36 ** (-0.40, -0.32)	-0.10 (-0.13, -0.06)		*-0.31 ** (-0.37, -0.26)	*-0.18 ~* (-0.24, -0.12)	*-0.11 ~* (-0.17, -0.06)	-0.09 (-0.14, -0.04)	-0.05 (-0.09, -0.01)
Edu PRI	0.03 (-0.03, 0.09)	*-0.11 ~* (-0.16, -0.06)		*-0.23 ~* (-0.29, -0.17)	*-0.15 ~* (-0.22, -0.08)	-0.08 (-0.15, -0.01)	*-0.14 ~* (-0.20, -0.09)	-0.09 (-0.15, -0.04)
Edu WO	-0.10 (-0.14, -0.05)	0.06 (0.02, 0.10)		*0.18 ~* (0.12, 0.24)	*0.18 ~* (0.12, 0.24)	0.03 (-0.02, 0.09)	*0.16 ~* (0.11, 0.21)	*0.24 ~* (0.19, 0.28)
Neutral Responding								
Edu WO	0.05 (0.01, 0.10)			*-0.14 ~* (-0.20, -0.09)	*-0.16 ~* (-0.23, -0.09)	*-0.18 ~* (-0.23, -0.13)		-0.04 (-0.09, -0.00)

~→ small effect; *→ medium effect; #→ large effect
[1]15-24 Years (Age 1524), 65 Years and Older (Age 65Ol).
[2]Primary Education (Edu PRI), WO (Edu WO).
[3]Family (FA), Health (HE), Housing (HO), Income (IN), Personality (PE), Politics (PO), Religion (RE), Work (WO).

Therefore, for each UAB and category, we considered the number of surveys for which at least a small either positive or negative effect was found. See Table 4.5. It is striking that relatively many cells or category-UAB pairs showed both positive and negative effects (marked by "2" in Table 4.5). This means that a category may show *more* of a specific UAB for some surveys, while *less* for other surveys. For instance, consider the category 15-24 years for the UAB answering "won't tell" (WT) in Table 4.5. Here, this age category showed more "won't tell"-answers than the other categories combined for one survey, while less "won't tell"-answers for another survey. For a more liberal perspective on consistency, we elaborate on the cases for which more than half of the applicable surveys showed either positive or

negative effect sizes (see Table 4.4). Strikingly, this is applicable to only seven out of the 96 possible cases (as we have results for eight UABs and twelve categories) and at a maximum of only 75% of the applicable surveys.

Table 4.5
The Categories for Age and Education (Edu) with either at Least Two Positive *or* Two Negative Effect Sizes Receiving a "1" (Unidirectional results) and the Categories with at Least One Positive *and* One Negative Effect Size Receiving a "2" (Contrasting results) for All Behaviours*

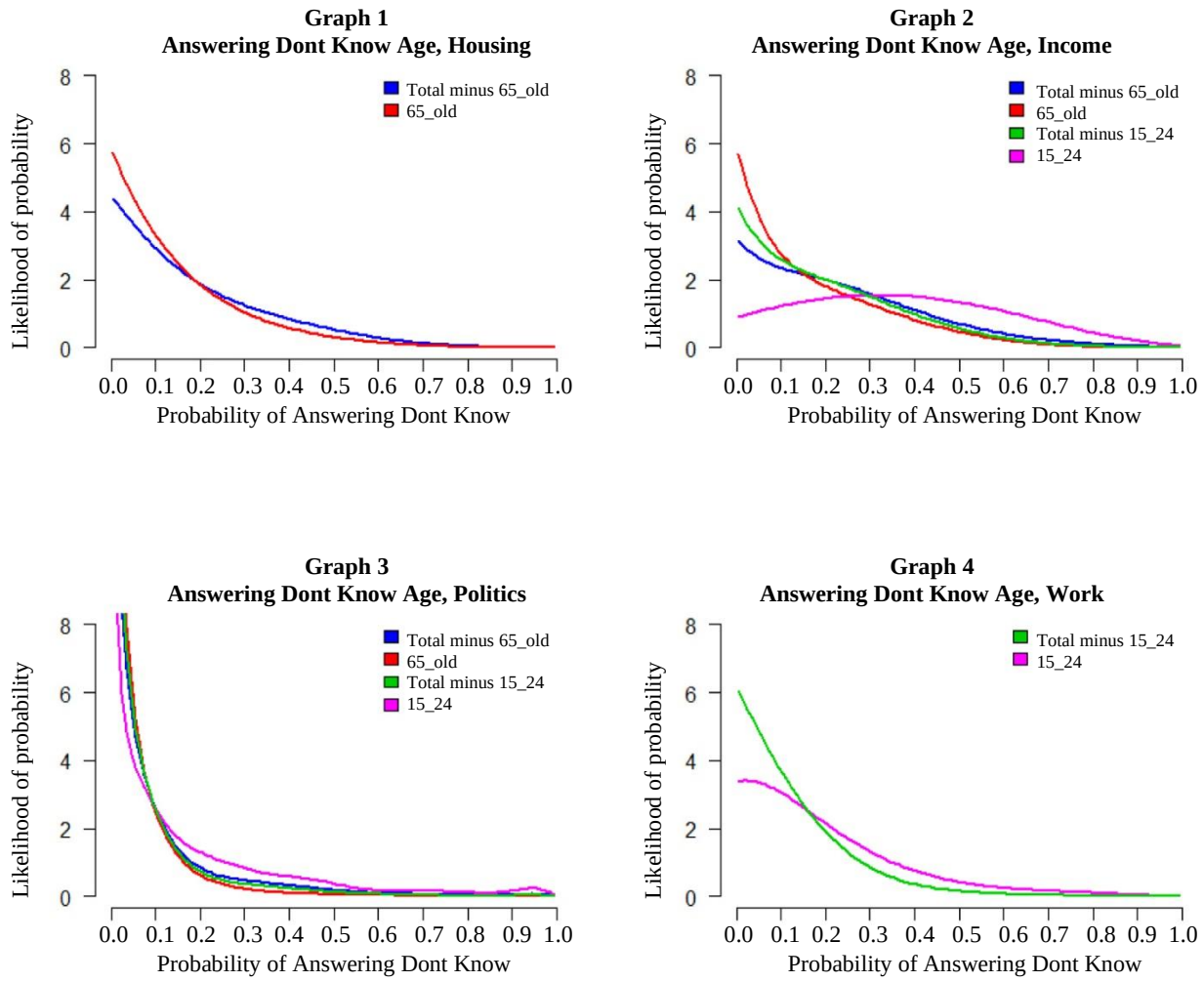
	Number of Surveys	**3**	**5**	**4/5**	**4/5**	**4/5/6**	**6/7**	**7**	**8**
	Answer Behaviour	**WT**	**AC**	**NE**	**EX**	**DK**	**ST**	**PR**	**SD**
Age	15-24 years	2				1	2	1	2
	25-34 years						2	2	2
	35-44 years						1	2	2
	45-54 years						1	1	
	55-64 years					1			2
	65 years or older					1	1	2	2
Edu	Primary education		1	1		1		1	2
	VMBO		1					2	2
	HAVWO								2
	MBO								
	HBO		1			1		1	1
	WO			1	2	1	1	1	2

The empty cells refer to either no effects, or one positive effect, or one negative effect.
*Answering Won't Tell (WT), Acquiescence (AC), Neutral Responding (NE), Extreme Responding (EX), Answering Don't Know (DK), Straightlining (ST), Primacy Responding (PR), Socially Desirable Responding (SD).

For the UAB answering "don't know", Table 4.4 shows that respondents 15-24 years of age gave more "don't know"-answers and respondents of 65 years or older gave less "don't know"-answers than other respondents for multiple surveys (see Graphs 1 through 4 in Figure 4.2). Respondents who finished only primary education gave more "don't know"-answers than other respondents for various surveys. For primacy responding, we found that respondents 15-24 years of age or who finished only primary education chose less early response options than other respondents for multiple surveys. Respondents who finished the highest educational level chose more early response options and less neutral responses than other respondents for various surveys.

In summary, the results refer to an absence of UAB consistency across all or almost all surveys: Both satisficing and sensitivity-based UABs did not emerge consistently across surveys. We conclude that respondents' UAB across surveys may be more influenced by the survey and its topic and items than solely by the age or educational level of the respondent. We close with a discussion in the following section.

Figure 4.2 Consistently *More* "Don't Know"-Answers for Respondents Aged 15-24 (purple) for the Surveys Income, Politics, and Work (see Graphs 2, 3, and 4 Respectively); Consistently *Less* "Don't Know"-Answers for Respondents Aged 65 or Older (red) for the Surveys Housing, Income, and Politics (see Graphs 1, 2, and 3 respectively).



5. Conclusion and discussion

In this study, we investigated to what extent cognitive ability is associated with a high occurrence of undesirable answer behaviour (UAB) *consistently* across different surveys. For cognitive ability, we used the respondent characteristics age and educational level. The occurrence of UAB is indicated by varying uncertainty, as every respondent filled out a different number of the items that were applicable to each behaviour. To take this varying uncertainty into account, we used an adaptation of the robust effect size statistic Cliff's Delta to compare groups of respondents in the form of density distributions or *respondent profiles*. The UAB of respondents from a specific category (for instance "15-24 years" for the characteristic "age") was compared to the UAB of respondents from the other categories of the

characteristic together. For our study, we included the specific satisficing behaviours "answering don't know", "acquiescence", "neutral responding", "extreme responding", "primacy responding", and "straightlining"; the specific sensitivity-based behaviours "socially desirable responding" and "answering won't tell"; and the respondent characteristics "age" and "education".

Considering all surveys together overall, specific satisficing and sensitivity-based behaviours are evident for specific age and educational groups. However, *there is no consistency across surveys present for the age and educational categories for any of the UABs*. This study used response data from a panel consisting of the same respondents. In general, if UAB consistency was to be expected at all, this should particularly be found in such a panel. If respondents would have any predisposition to show a behaviour style or pattern, this should especially occur while getting familiar with filling out multiple panel surveys within a specific time span. The fact that we did not find such patterns means that cognitive ability is most likely not a predictor of consistent UAB across surveys.

Considering consistency from a more liberal perspective, specific forms of satisficing across surveys seem evident for specific respondents in particular. Young and lower educated respondents gave relatively more "don't know"-answers; higher educated respondents chose relatively more answering options early in the list; young and lower educated respondents chose relatively less answering options early in the list; and higher educated respondents showed relatively less neutral responses for multiple surveys. However, there is no category for age or education that showed specific UAB consistently across *all or almost all* surveys.

Note that within a single survey, items are clustered around a central topic and may also be similar in their characteristics. This means that some item interdependency may occur within surveys. If we would have found consistent response patterns across surveys, these patterns may have been influenced by such item interdependency. Obviously, some respondents may be more sensitive to item interdependency in showing UAB across surveys than others. In our study, we did not find any consistent response patterns across surveys. This means that item interdependency was unlikely to exert a structurally different influence on the various categories of respondents across surveys.

Our results seem to go beyond the absence of UAB consistency across surveys. As the more surveys were applicable to an UAB, the more contrasting outcomes were found; many categories were associated with relatively *more* of an UAB for some surveys, while relatively *less* of that UAB for other surveys. Most contrasting results were found for giving socially desirable responses. More evidence was found for contrasting UAB than for consistent UAB across surveys. This evidence is not compatible to our idea that specific groups will show consistency for at least some of the specific UABs across most or all surveys. *Overall, we conclude that the occurrence of UAB cannot unambiguously be attributed to the respondent's cognitive ability, but may be substantially determined by the characteristics of the survey and its items instead.*

Following this conclusion, we do not recommend survey-independent adaptive survey design for respondents based on their cognitive ability. The findings for age and educational level are not consistent

and clearly differ depending on both survey and UAB. In essence, this means that our outcomes confirm the different associations and their different directions of the existing literature. The added value of our study is the overarching overview for age and educational level, systematically examined across a set of ten different surveys for a range of eight different UABs. We conclude that age and educational level may be taken into account for adaptive survey design only for specific surveys and survey topics.

In our study, we did not focus on UAB of *identified* individual or groups of respondents. For all age and educational categories, each respondent was considered for every applicable survey that the respondent participated in. Thus, for the consistency analysis of a category, some respondents were considered for only one or two surveys, while other respondents were considered for all or almost all surveys. Our purpose was neither to attribute UAB to individual or groups of identified respondents, nor to compare them between surveys for the same category and UAB. Considering respondents multiple times, for each applicable survey, was the strength of our study. Taking into account every respondent who fell into a category for every applicable survey resulted in large groups per survey. We compared respondent profiles of large groups for a single category to respondent profiles of large groups for the remaining categories. This means that we focussed on the association between the respondent's *characteristics* and potentially consistent UAB across surveys. In other words, we did not attribute UAB to identified respondents, but to the specific category (for instance respondents aged 15-24) in which they were placed. Considered from this approach, we note that we deliberately did not use a more classic method like cross-classified multilevel analysis (see for instance Olson and Smyth, 2015; Olson, Smyth and Ganshert, 2019) that takes into account repeated measurements of individual respondents. The focus of our study was placed on visualizing summaries of UAB and comparing subgroups that share the same characteristic.

We used the comparisons between a category and the remaining categories together for age and education to answer our consistency research question. For this purpose, we used an adaptation of Cliff's Delta; a robust effect size measure that was both useful because of its many advantages regarding our data, and sufficient for comparing two groups representing a specific category versus the remaining categories. In case of differences in expected group value or group shape, follow-up research may zoom in on these differences to reveal characteristics of subgroups showing relatively more of an UAB for specific surveys and their topics and items. Other relevant characteristics like respondent gender and origin may also be investigated. In particular, we would be interested in single groups with higher expected values than the other groups for a characteristic and in the respondents who are located to the right of the respondent profile.

Other follow-up research using the profile method may focus on the relation between *item characteristics* and UAB. Just as respondent characteristics, item characteristics have their influence on data quality and may be associated with measurement error. See Bais et al. (2019); Beukenhorst, Buelens, Engelen, Van der Laan, Meertens and Schouten (2014); Campanelli et al. (2011); Gallhofer, Scherpenzeel and Saris (2007), and Saris and Gallhofer (2007) for overviews of item characteristics and their relation to

measurement error. Items can be coded on the presence or absence of characteristics like for instance question sensitivity. Hence, items that are coded as sensitive could be compared to items that are not coded as sensitive on the occurrence of UAB. In this way, the presence of item characteristics may be connected to UAB for the items of whole surveys specifically or across the items of multiple surveys more generally. Based on such associations, an overview of present item characteristics and their relation to UAB and measurement error may be obtained.

Acknowledgements

We would like to thank Joost van der Neut for contributing to the adaptation of Cliff's Delta. We would like to thank CentERdata for the availability of LISS Panel data.

Appendix A

Table A.1
The behaviour occurrence proportions for respondents aged 65-74 (65+) and respondents aged 75 or older (75+) for all behaviours*, in total and for all surveys**

	TO	AS	FA	HE	HO	IN	PE	PO	RE	WO	LF
SD 65+	0.66	0.95	0.61	0.66	***	0.79	0.77	0.59	0.27	0.77	
SD 75+	0.65	0.96	0.60	0.64		0.78	0.76	0.58	0.30	0.79	
PR 65+	0.33		0.49	0.65		0.36	0.25	0.18	0.68	0.17	
PR 75+	0.31		0.50	0.65		0.33	0.24	0.16	0.66	0.13	
DK 65+	0.06				0.07	0.16		0.06	0.00		
DK 75+	0.06				0.07	0.14		0.07	0.00		
ST 65+	0.10		0.05	0.36		0.32	0.02	0.07	0.24		
ST 75+	0.08		0.04	0.25		0.29	0.01	0.06	0.19		
WT 65+	0.05				0.02	0.04					0.03
WT 75+	0.04				0.01	0.03					0.03
AC 65+	0.47		0.44				0.50	0.45	0.19		
AC 75+	0.49		0.42				0.51	0.48	0.21		
NE 65+	0.22		0.28			0.25	0.21	0.22			
NE 75+	0.21		0.28			0.25	0.21	0.22			
EX 65+	0.19		0.37			0.11	0.23	0.11			
EX 75+	0.20		0.40			0.11	0.25	0.10			

*Socially Desirable Responding (SD), Primacy Responding (PR), Answering "Don't Know" (DK), Straightlining (ST), Answering "Won't Tell" (WT), Acquiescence (AC), Neutral Responding (NE), Extreme Responding (EX).
**Total (TO), Assets (AS), Family (FA), Health (HE), Housing (HO), Income (IN), Personality (PE), Politics (PO), Religion (RE), Work (WO), Labour Force Survey (LF).
*** Note that empty cells refer either to surveys that were not applicable to the specific behaviour or to a situation in which one subgroup contained no or only a few respondents.

References

Aichholzer, J. (2013). Intra-individual variation of extreme response style in mixed-mode panel studies. *Social Science Research*, 42, 957-970. doi: http://dx.doi.org/10.1016/j.ssresearch.2013.01.002.

Alwin, D.F., and Krosnick, J.A. (1991). The reliability of survey attitude measurement. The influence of question and respondent attributes. *Sociological Methods and Research*, 20,139-181.

Andersen, H., and Mayerl, J. (2019). Responding to socially desirable and undesirable topics: Different types of response behaviour? *Methods, Data, Analyses: A Journal for Quantitative Methods and Survey Methodology (MDA)*, 13(1), 7-35. https://doi.org/10.12758/mda.2018.06.

Andrews, F.M., and Herzog, A.R. (1986). The quality of survey data as related to age of respondent. *Journal of the American Statistical Association*, 81(394), 403-410.

Antoni, M., Bela, D. and Vicari, B. (2019). Validating earnings in the German National Educational Panel Study: Determinants of measurement accuracy of survey questions on earnings. *Methods, Data, Analyses: A Journal for Quantitative Methods and Survey Methodology (MDA)*, 13(1), 59-90. https://doi.org/10.12758/mda.2018.08.

Bachman, J.G., and O'Malley, P.M. (1984). Yea-saying, nay-saying, and going to extremes: Black-white differences in response styles. *Public Opinion Quarterly*, 48, 491-509.

Bais, F. (2021). *Constructing Behaviour Profiles for Answer Behaviour Across Surveys*. Dissertation, Utrecht University. https://doi.org/10.33540/538.

Bais, F., Schouten, B., Lugtig, P., Toepoel, V., Arends-Tóth, J., Douhou, S., Kieruj, N., Morren, M. and Vis, C. (2019). Can survey item characteristics relevant to measurement error be coded reliably? A case study on eleven Dutch General Population Surveys. *Sociological Methods and Research*, 48(2), 263-295. https://doi.org/10.1177/0049124117729692.

Bais, F., Schouten, B. and Toepoel, V. (2020). Investigating response patterns across surveys: Do respondents show consistency in undesirable answer behaviour over multiple surveys? *Bulletin de Méthodologie Sociologique*, 147-148(1-2), 150-168. https://doi.org/10.1177/0759106320939891.

Beatty, P., and Herrmann, D. (2002). To answer or not to answer: Decision processes related to survey item nonresponse. In *Survey Nonresponse*, (Eds., R.M. Groves, D.A. Dillman, J.L. Eltinge and R.J.A. Little), First Edition, New York: John Wiley & Sons, Inc., 71-86.

Beukenhorst, D., Buelens, B., Engelen, F.,Van der Laan, J., Meertens, V. and Schouten, B. (2014). *The Impact of Survey Item Characteristics on Mode-Specific Measurement Bias in the Crime Victimisation Survey*. CBS Discussion paper 2014-16. Statistics Netherlands, The Hague.

Billiet, J.B., and McClendon, J.M. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling*, 7, 608-628. doi: http://dx.doi.org/10.1207/S15328007SEM0704_5.

Binswanger, J., Schunk, D. and Toepoel, V. (2013). Panel conditioning in difficult attitudinal questions. *Public Opinion Quarterly*, 77, 783-797.

Bishop, G.F., Tuchfarber, A.J. and Oldendick, R.W. (1986). Opinions on fictitious issues: The pressure to answer survey questions. *Public Opinion Quarterly*, 50, 240-250.

Bradburn, N., Sudman, S., Blair, E. and Stocking, C. (1978). Question threat and response bias. *Public Opinion Quarterly*, 42, 221-234.

Campanelli, P., Nicolaas, G., Jäckle, A., Lynn, P., Hope, S., Blake, M. and Gray, M. (2011). *A Classification of Question Characteristics Relevant to Measurement (Error) and Consequently Important for Mixed Mode Questionnaire Design*. Paper presented at the Royal Statistical Society, October 11, London, UK.

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, 114, 494-509.

Cliff, N. (1996a). Answering ordinal questions with ordinal data using ordinal statistics. *Multivariate Behavioral Research*, 31, 331-350.

Cliff, N. (1996b). *Ordinal Methods for Behavioral Data Analysis*. New Jersey: Lawrence Erlbaum Associates.

Dekking, F.M., Kraaikamp, C., Lopuhaä, H.P. and Meester, L.E. (2005). *A Modern Introduction to Probability and Statistics – Understanding Why and How*. Springer. https://doi.org/10.1007/1-84628-168-7.

De Leeuw, E.D. (1992). *Data Quality in Mail, Telephone, and Face-to-Face Surveys*. Amsterdam: TT-Publicaties.

De Leeuw, E.D., Hox, J.J. and Dillman, D. (2008). *International Handbook of Survey Methodology*. Taylor & Francis Group.

DeMaio, T.J. (1984). Social desirability and survey measurement: A review. In *Surveying Subjective Phenomena*, (Eds., C.F. Turner and E. Martin). New York: Russell Sage Foundation, 2, 257-281.

Díaz de Rada, V., and Domínguez, J.A. (2015). The quality of responses to grid questions as used in web questionnaires (compared with paper questionnaires). *International Journal of Social Research Methodology*, 18, 337-348. doi: http://dx.doi.org/10.1080/13645579.2014.895289.

Fricker, S., Galesic, M., Tourangeau, R. and Yan, T. (2005). An experimental comparison of web and telephone surveys. *Public Opinion Quarterly*, 69, 370-392. doi: http://dx.doi.org/10.1093/poq/nfi027.

Galesic, M., Tourangeau, R., Couper, M.P. and Conrad, F.G. (2008). Eye-tracking data new insights on response order effects and other cognitive shortcuts in survey responding. *Public Opinion Quarterly*, 72, 892-913.

Gallhofer, I.N., Scherpenzeel, A. and Saris, W.E. (2007). *The Code-Book for the SQP Program*, available at http://www.europeansocialsurvey.org/methodology/.

Goedhart, J. (2016). *Calculation of a Distribution Free Estimate of Effect Size and Confidence Intervals Using VBA/Excel*. doi: http://dx.doi.org/10.1101/073999.

Greenleaf, E.A. (1992). Measuring extreme response style. *Public Opinion Quarterly*, 56, 328-351. http://www.jstor.org/stable/2749156.

He, J., and Van de Vijver, F.J.R. (2013). A general response style factor: Evidence from amulti-ethnic study in the Netherlands. *Personality and Individual Differences*, 55, 794-800. http://dx.doi.org/10.1016/j.paid.2013.06.017.

He, J., Van de Vijver, F.J.R., Espinosa, A.D. and Mui, P.H. (2014). Toward a unification of acquiescent, extreme, and midpoint response styles: A multilevel study. *International Journal of Cross Cultural Management*, 14, 306-322. doi: http://dx.doi.org/10.1177/1470595814541424.

Heerwegh, D., and Loosveldt, G. (2011). Assessing mode effects in a national crime victimization survey using structural equation models: Social desirability bias and acquiescence. *Journal of Official Statistics*, 27, 49-63.

Hess, M.R., and Kromrey, J.D. (2004). *Robust Confidence Intervals for Effect Sizes: A Comparative Study of Cohen's d and Cliff's Delta under Non-Normality and Heterogeneous Variances*. Paper Presented at the Annual Meeting of the American Educational Research Association, San Diego, California.

Holbrook, A.L., Green, M.C. and Krosnick, J.A. (2003). Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. *Public Opinion Quarterly*, 67, 79-125.

Hox, J.J., De Leeuw, E. and Kreft, I.G. (1991). The effect of interviewer and respondent characteristics on the quality of survey data: A multilevel model. In *Measurement Errors in Surveys*, (Eds., P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz and S. Sudman), New York: John Wiley & Sons, Inc., 439-461.

Jann, B., Krumpal, I. and Wolter, F. (2019). Editorial: Social desirability bias in surveys – Collecting and analyzing sensitive data. *Methods, Data, Analyses: A Journal for Quantitative Methods and Survey Methodology (MDA)*, 13(1), 3-6.

Johnson, T., and Van de Vijver, F.J.R. (2003). Social desirability in cross cultural research. In *Cross-Cultural Survey Methods*, (Eds., J. Harness, F.J.R. van de Vijver and P. Mohler.), New York: John Wiley & Sons, Inc., 193-202.

Kalton, G., Roberts, J. and Holt, D. (1980). The effects of offering a middle response option with opinion questions. *Statistician*, 29, 65-78. http://www.jstor.org/stable/2987495.

Kaminska, O., McCutcheon, A. and Billiet, J. (2010). Satisficing among reluctant respondents in a cross-national context. *Public Opinion Quarterly*, 74, 880-906. doi: http://dx.doi.org/10.1093/poq/nfq062.

Kieruj, N.D., and Moors, G. (2013). Response style behavior: Question format dependent or personal Style? *Quality and Quantity*, 47, 193-211. doi: http://dx.doi.org/10.1007/s11135-011-9511-4.

Kreuter, F., Presser, S. and Tourangeau, R. (2008). Social desirability bias in CATI, IVR, and web surveys: The effects of mode and question sensitivity. *Public Opinion Quarterly*, 72, 847-865. doi: http://dx.doi.org/10.1093/poq/nfn063.

Krosnick, J.A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5, 213-236.

Krosnick, J.A. (1992). The impact of cognitive sophistication and attitude importance onresponse order effects and question order effects. In *Order Effects in Social and Psychological Research*, (Eds., N. Schwarz and S. Sudman), New York: Springer, 203-218.

Krosnick, J.A. (1999). Survey research. *Annual Review of Psychology*, 50, 537-567.

Krosnick, J.A., and Alwin, D.F. (1987). An evaluation of a cognitive theory of response order effects in survey measurement. *Public Opinion Quarterly*, 51, 201-219.

Krosnick, J.A., and Alwin, D.F. (1989). Aging and susceptibility to attitude change. *Journal of Personality and Social Psychology*, 57, 416-425.

Krosnick, J.A., and Fabrigar, L.R. (1997). Designing rating scales for effective measurement in surveys. In *Survey Measurement and Process Quality*, (Eds., L. Lyberg, P. Biemer, M. Collins, L. Decker, E. de Leeuw, C. Dippo, N. Schwarz and D. Trewin), New York: John Wiley & Sons, Inc., 141-164.

Krosnick, J.A., Holbrook, A.L., Berent, M.K., Carson, R.T., Hanemann, W.M., Kopp, R.J., Mitchell, R.C., Presser, S., Ruud, P.A., Smith, V.K., Moody, W.R., Green, M.C. and Conaway, M. (2002). The impact of “no opinion” response options on data quality: Non-attitude reduction or an invitation to satisfice? *Public Opinion Quarterly*, 66, 371-403.

Krosnick, J.A., Narayan, S. and Smith, W.R. (1996). Satisficing in surveys: Initial evidence. *New Directions for Evaluation*, 70, 29-44.

Leigh, J.H., and Martin, C.R. (1987). Don't know item nonresponse in a telephone survey: Effects of question form and respondent characteristics. *Journal of Marketing Research*, 24, 418-424.

Lensvelt-Mulders, G.J.L.M. (2008). Surveying sensitive topics. In *International Handbook of Survey Methodology*, (Eds., E.D. de Leeuw, J.J. Hox and D.A. Dillman). New York: Taylor and Francis, Psychology Press, EAM series, 461-478.

Marín, G., Gamba, R.J. and Marín, B.V. (1992). Extreme response style and acquiescence among hispanics. *Journal of Cross-Cultural Psychology*, 23, 498-509.

McClendon, M.J. (1986). Response-order effects for dichotomous questions. *Social Science Quarterly*, 67, 205-211.

McClendon, M.J. (1991). Acquiescence and recency response-order effects in interview surveys. *Sociological Methods and Research*, 20, 60-103.

Medway, R., and Tourangeau, R. (2015). Response quality in telephone surveys. Do pre-paid cash incentives make a difference? *Public Opinion Quarterly*, 79, 524-543. doi: http://dx.doi.org/10.1093/poq/nfv011.

Meisenberg, G., and Williams, A. (2008). Are acquiescent and extreme response styles related to low intelligence and education? *Personality and Individual Differences*, 44, 1539-1550. https://doi.org/10.1016/j.paid.2008.01.010.

Messick, S.J. (1966). The psychology of acquiescence: An interpretation of research evidence. In *Response Set in Personality Assessment*, (Ed., I.A. Berg), Chicago: Aldine, 115-145.

Narayan, S., and Krosnick, J.A. (1996). Education moderates some response effects in attitude measurement. *Public Opinion Quarterly*, 60, 58-88.

Olson, K., and Smyth, J.D. (2015). The effect of CATI questions, respondents, and interviewers on response time. *Journal of Survey Statistics and Methodology*, 3, 361-396. doi: http://dx.doi.org/10.1093/jssam/smv021.

Olson, K., Smyth, J.D., and Ganshert, A. (2019). The effects of respondent and question characteristics on respondent answering behaviors in telephone interviews. *Journal of Survey Statistics and Methodology*, 7(2), 275- 308. https://doi.org/10.1093/jssam/smy006.

O'Muircheartaigh, C., Krosnick, J.A. and Helic, A. (2000). *Middle Alternatives, Acquiescence, and the Quality of Questionnaire Data*. Retrieved, October 1, 2009, from http://harrisschool.uchicago.edu/About/publications.

Paulhus, D.L. (2002). Socially desirable responding: The evolution of a construct. In *The Role of Constructs in Psychological and Educational Measurement*, (Eds., H.I. Braun, D.N. Jackson and D.E. Wiley), Mahwah, NJ: Erlbaum, 49-69.

Pickery, J., and Loosveldt, G. (1998). The impact of respondent and interviewer characteristics on the number of "No opinion" answers. A multilevel model for count data. *Quality and Quantity*, 32, 31-45.

Roberts, C. (2007). Mixing modes of data collection in surveys: A methodological review. ESRC National Centre for Research Methods, NCRM Methods Review Paper 008, UK. Retrieved July 2019 from http://eprints.ncrm.ac.uk/418/1/MethodsReviewPaperNCRM-008.pdf.

Roberts, C., and Jäckle, A. (2012). *Causes of Mode Effects: Separating out Interviewer and Stimulus Effects in Comparisons of Face-to-Face and Telephone Surveys*. ISER Working Paper, 2012-27. Colchester: University of Essex.

Roßmann, J., Gummer, T. and Silber, H. (2017). Mitigating satisficing in cognitively demanding grid questions: Evidence from two web-based experiments. *Journal of Survey Statistics and Methodology*.

Rousselet, G.A., Foxe, J.J. and Bolam, J.P. (2016). A few simple steps to improve the description of group results in neuroscience. *European Journal of Neuroscience*, 44, 2647-2651. doi: https://doi.org/10.1111/ejn.13400.

Rousselet, G.A., Pernet, C.R. and Wilcox, R.R. (2017). Beyond differences in means:robust graphical methods to compare two groups in neuroscience. *European Journal of Neuroscience*, 1-27. doi: http://dx.doi.org/10.1101/121079.

Saris, W.E., and Gallhofer, I.N. (2007). Estimation of the effects of measurement characteristics on the quality of survey questions. *Survey Research Methods*, 1(1), 29-43. https://doi.org/10.18148/srm/2007.v1i1.49.

Saris, W.E., Revilla, M., Krosnick, J.A. and Shaeffer, E. (2010). Comparing questions with agree/disagree response options to questions with item-specific response options. *Survey Research Methods*, 4(1), 61-79. http://www.surveymethods.org.

Schaeffer, N.C., and Presser, S. (2003). The science of asking questions. *Annual Review of Sociology*, 29, 65-88. doi: https://doi.org/10.1146/annurev.soc.29.110702.110112.

Schonlau, M., and Toepoel, V. (2015). Straightlining in web survey panels over time. *Survey Research Methods*, 9, 125-137. doi: https://doi.org/10.18148/srm/2015.v9i2.6128.

Schouten, B., and Calinescu, M. (2013). Paradata as input to monitoring representativeness and measurement profiles: A case study of the Dutch Labour Force Survey. In *Improving Surveys with Paradata: Analytic Uses of Process Information*, (Ed., F. Kreuter), Hoboken, NJ: Wiley, 231-258.

Schuman, H., and Presser, S. (1981). *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording and Context*. New York: Academic Press.

Shoemaker, P.J., Eichholz, M. and Skewes, E.A. (2002). Item nonresponse: Distinguishing between don't know and refuse. *International Journal of Public Opinion Research*, 14, 193-201.

Si, S.X., and Cullen, J.B. (1998). Response categories and potential cultural bias: Effects of an explicit middle point in cross-cultural surveys. *International Journal of Organizational Analysis*, 6, 218-230.

Stern, M.J., Dillman, D.A. and Smyth, J.D. (2007). Visual design, order effects, and respondent characteristics in a self-administered survey. *Survey Research Methods*, 1, 121-138. http://www.surveymethods.org.

Stricker, L.J. (1963). Acquiescence and social desirability response styles, item characteristics, and conformity. *Psychological Reports*, 12, 319-341.

Tarnai, J., and Dillman, D.A. (1992). Questionnaire context as a source of response differences in mail versus telephone surveys. In *Context Effects in Social and Psychological Research*, (Eds., N. Schwarz and S. Sudman), New York: Springer Verlag.

Tourangeau, R., Rips, L.J. and Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge: Cambridge University Press.

Tourangeau, R., and Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, 133, 859-883. doi: https://doi.org/10.1037/0033-2909.133.5.859.

Van Herk, H., Poortinga, Y.H. and Verhallen, T.M.M. (2004). Response styles in rating scales: Evidence of method bias in data from six EU countries. *Journal of Cross-Cultural Psychology*, 35, 346-360.

Van Rosmalen, J., Van Herk, H. and Groenen, P.J.F. (2010). Identifying response styles: A latent-class bilinear multinomial logit model. *Journal of Marketing Research*, 47(1), 157-172. doi: https://doi.org/10.1509/jmkr.47.1.157.

Vargha, A., and Delaney, H.D. (2000). A critique and improvement of the CL Common Language effect size statistics of McGraw and Wong. *Journal of Educational Behavioral Statistics*, 25(2), 101-132. doi: http://dx.doi.org/10.2307/1165329.

Vis-Visschers, R., Arends-Tóth, J., Giesen, D. and Meertens, V. (2008). *Het Aanbieden Van "Weet Niet" en Toelichtingen in Een Webvragenlijst*. Report DMH-2008-02-21-RVCS, Statistics Netherlands, Methodology Department, Heerlen, The Netherlands.

Ye, C., Fulton, J. and Tourangeau, R. (2011). More positive or more extreme? A meta-analysis of mode differences in response choice. *Public Opinion Quarterly*, 75(2), 349-365. doi: https://doi.org/10.1093/poq/nfr009.

Zhang, C. (2013). *Satisficing in Web Surveys: Implications for Data Quality and Strategies for Reduction*, (Ph.D.) Ann Arbor, MI: University of Michigan. Retrieved from http://deepblue.lib.umich.edu/handle/2027.42/97990.

Zhang, C., and Conrad, F.G. (2013). Speeding in Web surveys: The tendency to answer very fast and its association with straightlining. *Survey Research Methods*, 8, 127-135.

Survey Methodology, June 2022
Vol. 48, No. 1, pp. 225-249
Statistics Canada, Catalogue No. 12-001-X

A simulated annealing algorithm for joint stratification and sample allocation

Mervyn O'Luing, Steven Prestwich and S. Armagan Tarim[1]

Abstract

This study combines simulated annealing with delta evaluation to solve the joint stratification and sample allocation problem. In this problem, atomic strata are partitioned into mutually exclusive and collectively exhaustive strata. Each partition of atomic strata is a possible solution to the stratification problem, the quality of which is measured by its cost. The Bell number of possible solutions is enormous, for even a moderate number of atomic strata, and an additional layer of complexity is added with the evaluation time of each solution. Many larger scale combinatorial optimisation problems cannot be solved to optimality, because the search for an optimum solution requires a prohibitive amount of computation time. A number of local search heuristic algorithms have been designed for this problem but these can become trapped in local minima preventing any further improvements. We add, to the existing suite of local search algorithms, a simulated annealing algorithm that allows for an escape from local minima and uses delta evaluation to exploit the similarity between consecutive solutions, and thereby reduces the evaluation time. We compared the simulated annealing algorithm with two recent algorithms. In both cases, the simulated annealing algorithm attained a solution of comparable quality in considerably less computation time.

Key Words: Simulated annealing algorithm; Optimal stratification; Sample allocation; R software.

1. Introduction

In stratified simple random sampling, a population is partitioned into mutually exclusive and collectively exhaustive strata, and then sampling units from each of those strata are randomly selected. The purposes for stratification are discussed in Cochran (1977). If the intra-strata variances were minimized then precision would be improved. It follows that the resulting small samples from each stratum can be combined to give a small sample size.

To this end, we intend to construct strata which are internally homogeneous but which also accommodate outlying measurements. To do so, we adopt an approach which entails searching for the optimum partitioning of *atomic strata* (however, the methodology can also be applied to *continuous* strata) created from the Cartesian product of categorical stratification variables, see Benedetti, Espa and Lafratta (2008); Ballin and Barcaroli (2013, 2020).

The Bell number, representing the number of possible partitions (stratifications) of a set of atomic strata, grows very rapidly with the number of atomic strata (Ballin and Barcaroli, 2013). In fact, there comes a point where, even for a moderate number of atomic strata and the most powerful computers, the problem is intractable, i.e. there are no known efficient algorithms to solve the problem.

Many large scale combinatorial optimisation problems of this type cannot be solved to optimality, because the search for an optimum solution requires a prohibitive amount of computation time. This

1. Mervyn O'Luing, Insight Centre for Data Analytics, Department of Computer Science, University College Cork, Ireland. E-mail: mervyn.oluing@insight-centre.org; Steven Prestwich, Insight Centre for Data Analytics, Department of Computer Science, University College Cork, Ireland. E-mail: steven.prestwich@insight-centre.org; S. Armagan Tarim, Cork University Business School, University College Cork, Ireland. E-mail: armagan.tarim@ucc.ie.

compels one to use *approximisation algorithms* or *heuristics* which do not guarantee optimal solutions, but can provide approximate solutions in an acceptable time interval. In this way, one trades off the quality of the final solution against computation time (Van Laarhoven and Aarts, 1987). In other words, heuristic algorithms are developed to find a solution that is "good enough" in a computing time that is "small enough" (Sörensen and Glover, 2013).

A number of heuristic algorithms have been developed to search for optimal or near optimal solutions, for both univariate and multivariate scenarios of this problem. This includes the hierarchichal algorithm proposed by Benedetti et al. (2008), the genetic algorithm proposed by Ballin and Barcaroli (2013) and the grouping genetic algorithm proposed by O'Luing, Prestwich and Tarim (2019). Although effective, the evaluation function in these algorithms can be costly in terms of running time.

We add to this work with a simulated annealing algorithm (SAA) (Kirkpatrick, Gelatt and Vecchi, 1983; Černỳ, 1985). SAAs have been found to work well in problems such as this, where there are many local minima and finding an approximate global solution in a fixed amount of computation time is more desirable than finding a precise local minimum (Takeang and Aurasopon, 2019). We present a SAA to which we have added delta evaluation (see Section 5) to take advantage of the similarity between consecutive solutions and help speed up computation times.

We compared the performance of the SAA on atomic strata with that of the grouping genetic algorithm (GGA) in the *SamplingStrata* package (Ballin and Barcaroli, 2020). This algorithm implements the grouping operators described by O'Luing et al. (2019). To do this, we used sampling frames of varying sizes containing what we assume to be completely representative details for target and auxiliary variable columns.

Further to the suggestion of a *Survey Methodology* reviewer, we subsequently compared the SAA with a traditional genetic algorithm (TGA) used by Ballin and Barcaroli (2020) on continuous strata. In both sets of experiments, we used an initial solution created by the k-means algorithm (Hartigan and Wong, 1979) in a two-stage process (see Section 2.3 for more details).

Section 2 provides background information on atomic strata, introduces the SAA and motivates the addition of delta evaluation as a means to improve computation time. Two-stage simulated annealing is also discussed. Section 3 of the paper describes the cost function and evaluation algorithm. Section 4 provides an outline of the SAA. Section 5 presents the improved SAA with delta evaluation. Section 6 provides a comparison of the performance of the SAA with the GGA using an initial solution and fine-tuned hyperparameters. Section 7 then provides details of the comparison of the SAA with the genetic algorithm in Ballin and Barcaroli (2020) on continuous strata. Section 8 presents the conclusions and Section 9 suggests some further work. The Appendix contains background details on precision constraints, the hyperparameters, and the process of fine-tuning the hyperparameters for both comparisons as well as the computer specifications.

2. Background information

2.1 Stratification of atomic strata

Atomic strata are created using categorical auxiliary variable columns such as *age group*, *gender* or *ethnicity* for a survey of people or *industry*, *type of business* and *employee size* for business surveys. The cross-classification of the class-intervals of the auxiliary variable columns form the atomic strata.

Auxiliary variable columns which are correlated to the target variable columns may provide a gain in sample precision or *similarity*. Each target variable column, y_g, contains the value of the survey characteristic of interest, e.g. *total income*, for each population element in the sample.

Once these are created, we obtain summary statistics, such as the number, mean and standard deviation of the relevant observed values, from the one or more target variable columns that fall within each atomic stratum. The summary information is then aggregated in order to calculate the means and variances for each stratum which in turn are used to calculate the sample allocation for a given stratification.

The partitioning of atomic strata that provides the *global minimum* sample allocation, i.e. the minimum of all possible sample allocations for the set of possible stratifications, is known as an *optimal stratification*. There could be a multiple of such partitionings. Although an optimum stratification is *the* solution to the problem, each stratification represents a solution of varying quality (the lower the cost (*minimum* or *optimal* sample allocation) the higher the quality). For each stratification, the cost is estimated by the Bethel-Chromy algorithm (Bethel, 1985, 1989; Chromy, 1987). A more detailed description, and discussion of the methodology for this approach for joint determination of stratification and sample allocation, can be found in Ballin and Barcaroli (2013).

2.2 Simulated annealing algorithms

The basic principle of the SAA (Kirkpatrick et al., 1983; Černỳ, 1985) is that it can accept solutions that are inferior to the current best solution in order to find the global minima (or maxima). It is one of several stochastic local search algorithms, which focus their attention within a local neighbourhood of a given initial solution (Cortez, 2014), and use different stochastic techniques to escape from attractive local minima (Hoos and Stützle, 2004).

Based on physical annealing in metallurgy, the SAA is designed to simulate the controlled cooling process from liquid metal to a solid state (Luke, 2013). This controlled cooling uses the temperature parameter to compute the probability of accepting inferior solutions (Cortez, 2014). This acceptance probability is not only a function of the temperature, but also the difference in cost between the new solution and the current best solution. For the same difference in cost, a higher temperature means a higher probability of accepting inferior solutions.

For a given temperature, solutions are iteratively generated by applying a small, randomly generated, perturbation to the current best solution. Generally, in SAAs, a perturbation is the small displacement of a

randomly chosen particle (Van Laarhoven and Aarts, 1987). In the context of our problem, we take perturbation to mean the displacement (or re-positioning) of q (generally $q=1$) randomly chosen atomic strata from one randomly chosen stratum to another.

With a perturbation, the current best solution transitions to a new solution. If a perturbation results in a lower cost for the new solution, or if there is no change in cost, then that solution is always selected as the current best solution. If the new solution results in a higher cost, then it is accepted at the above mentioned acceptance probability. This acceptance condition is called the Metropolis criterion (Metropolis, Rosenbluth, Rosenbluth, Teller and Teller, 1953). This process continues until the end of the sequence, at which point the temperature is decremented and a new sequence begins.

If the perturbations are minor, then the current solution and the new solution will be very similar. Indeed, in our SAA we are assuming only a slight difference between consecutive solutions owing to such perturbations (see Section 4.1 for more details). For this reason we have added delta evaluation, which will be discussed further in Section 5, to take advantage of this similarity and help improve computation times.

Accordingly, and as mentioned in the introduction, we present a SAA with delta evaluation and compare it with the GGA when both are combined with an initial solution. We also compare it with a genetic algorithm used by Ballin and Barcaroli (2020) on continuous strata. We provide more background details on initial solutions in Section 2.3 below.

2.3 Two-stage simulated annealing

A two stage simulated annealing process, where an initial solution is generated by a heuristic algorithm in the first stage, has been proposed for problems such as the *cell placement problem* (Grover, 1987; Rose, Snelgrove and Vranesic, 1988) or the *graph partitioning problem* (Johnson, Aragon, McGeoch and Schevon, 1989). Lisic, Sang, Zhu and Zimmer (2018) combined an initial solution, generated by the k-means algorithm, with a simulated annealing algorithm, for a problem similar in nature to this problem, but where the sample allocation as well as strata number are fixed, and the algorithm searches for the optimal arrangement of sampling units between strata.

The simulated annealing algorithm used by Lisic et al. (2018) starts with an initial solution (stratification and sample allocation to each stratum) and, for each iteration, generates a new candidate solution by moving one atomic stratum from one stratum to another and adjusting the sample allocation for that stratification. Each candidate solution is then evaluated to measure the coefficient of variation (CV) of the target variables and is accepted, as the new current best solution, if its objective function is less than the preceding solution. Inferior quality solutions are also accepted at a probability, ρ, which is a function of a tunable temperature parameter and the change in solution quality between iterations. The temperature cools, at a rate which is also tunable, as the number of iterations increases.

Following this work, Ballin and Barcaroli (2020) recommended combining an initial solution, generated by k-means, with the grouping and traditional genetic algorithms. They demonstrate that the k-means algorithm provides better starting solutions when compared with the starting solution generated by a stochastic approach. We also combine a k-means initial solution with the SAA in the experiments described in Sections 6 and 7.

3. The joint stratification and sample allocation problem

Our aim is to partition L atomic strata into H non-empty sub-populations or strata. A partitioning represents a stratification of the population. We aim to minimise the sample allocation to this stratification while keeping the measure of similarity less than or equal to the upper limit of precision, ε_g. This similarity is measured by the CV of the estimated population total for each one of G target variable columns, $\hat{T}_g$. We indicate by n_h the sample allocated to stratum h and the survey cost for a given stratification is calculated as follows:

$$C\left(n_1, \ldots, n_H\right) = \sum_{h=1}^{H} C_h n_h$$

where C_h is the average cost of surveying one unit in stratum h and n_h is the sample allocation to stratum h. In our analysis C_h is set to 1.

The variance of the estimator is given by:

$$\mathrm{VAR}\left(\hat{T}_g\right) = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_{h,g}^2}{n_h} \qquad \left(g = 1, \ldots, G\right)$$

where N_h is the number of units in stratum h and $S_{h,g}^2$ is the variance of stratum h for each target variable column g.

As mentioned above ε_g is the upper precision limit for the CV for each $\hat{T}_g$:

$$\mathrm{CV}\left(\hat{T}_g\right) = \frac{\sqrt{\mathrm{VAR}\left(\hat{T}_g\right)}}{E\left(\hat{T}_g\right)} \leq \varepsilon_g.$$

The problem can be summarised in this way:

$$\begin{aligned} &\min && n = \sum_{h=1}^{H} n_h \\ &\text{subject to} && \mathrm{CV}\left(\hat{T}_g\right) \leq \varepsilon_g \qquad \left(g = 1, \ldots, G\right). \end{aligned}$$

To solve the allocation problem for a particular stratification with the Bethel-Chromy algorithm the upper precision constraint for variable g can be expressed as follows:

$$\begin{aligned}\mathrm{CV}\left(\hat{T}_g\right)^2 \le \varepsilon_g^2 \quad &\equiv \sum_{h=1}^{H} \frac{N_h^2 S_{h,g}^2}{n_h} - N_h S_{h,g}^2 \le E\left(\hat{T}_g^2\right)\varepsilon_g^2 \\ &\equiv \sum_{h=1}^{H} \frac{N_h^2 S_{h,g}^2}{\left(E\left(\hat{T}_g^2\right)\varepsilon_g^2 + \sum_{h=1}^{H} N_h S_{h,g}^2\right) n_h} \le 1.\end{aligned}$$

Then we substitute

$$\frac{N_h^2 S_{h,g}^2}{\left(E\left(\hat{T}_g^2\right)\varepsilon_g^2 + \sum_{h=1}^{H} N_h S_{h,g}^2\right)}$$

with ξ_h, g and replace the problem summary with the following:

$$\begin{aligned}&\min\ n = \sum_{h=1}^{H} n_h \\ &\sum_{h=1}^{H} \frac{\xi_h, g}{n_h} \le 1 \quad \left(g = 1, \ldots, G\right)\end{aligned}$$

where $\frac{1}{n_h} > 0.$ The Bethel-Chromy algorithm uses Lagrangian multipliers to derive a solution for each n_h.

$$\frac{1}{n_h} = \begin{cases} \dfrac{\sqrt{1}}{\left(\sqrt{\sum_{g=1}^{G} \alpha_g \xi_h, g}\ \sum_{h=1}^{H} \sqrt{\sum_{g=1}^{G} \alpha_g \xi_h, g}\right)} & \text{if } \sum_{g=1}^{G} \alpha_g \xi_h, g > 0 \\ +\infty & \text{otherwise} \end{cases}$$

where

$$\alpha_g = \frac{\lambda_g}{\sum_{g=1}^{G} \lambda_g},$$

and λ_g is the Lagrangian multiplier (Benedetti et al., 2008). The algorithm starts with a default setting for each α_g and uses gradient descent to converge to a final value for them.

4. Outline of the simulated annealing algorithm

The SAA with delta evaluation is described in Algorithm 1 below. We then describe the heuristics we have used in the SAA. Delta evaluation is explained in more detail in Section 5.

Algorithm 1 Simulated annealing algorithm

Function SIMULATEDANNEALING (S is the starting solution, f is the evaluation function (Bethel-Chromy algorithm), *best* is the current best solution, *BSFSF* is the best solution found so far, *maxit* is the maximum number of sequences, J is the length of sequence, T_{max} is the starting temperature, T_{min} is the minimum temperature, *DC* is the Decrement Constant, $L_{max\%}$ is a % of L (number of atomic strata), $P(H+1)$ is the probability of a new stratum, $H+1$, being added)

```
    T ← T_max
    best ← S
    Cost(best) ← f(best)                                  ►using Bethel-Chromy algorithm
    while i < maxit && T > T_min do
        if RANDOM(0,1) ≤ 1/J then
            for l = 1 to L do
                if RANDOM(0,1) ≤ P(H+1) then
                    move atomic stratum l to new stratum H+1      ►see Section 4.3
                end if
            end for
        end if
        for j = 1 to J do
            if i = 1 & j = 1 then q = L × L_max%
            else if i = 1 & j > 1 then q = ceiling(q × 0.99)      ►0.99 is not tunable
            else if i > 1 then q = 1
            end if
            Randomly select h and h′
            next ← PERTURBATION(best)
                                                          ►Assign q atomic strata from h and h′
            Cost(next) ← f(best)                          ►using delta evaluation
            ΔE ← COST(next) − COST(best)
            if ΔE ≤ 0 then
                best ← next
            else if RANDOM(0,1) < e^(−ΔE/T) then          ►Metropolis Criterion
                best ← next
            end if
            if best <= BSFSF then
                BSFSF ← best
            end if
        end for
        T ← T * DC
    end while
    return BSFSF
end function
```

4.1 Perturbation

Consider the following solution represented by the stratification:

$$\{1,3\},\ \{2\},\ \{4,5,6\}.$$

The integers within each stratum represent atomic strata. In perturbation, the new solution below is created by arbitrarily moving atomic strata, in this example $q=1$, from one randomly chosen stratum to another.

$$\{1,3,2\},\{\varnothing\},\{4,5,6\}.$$

The first stratum gains an additional atomic stratum $\{2\}$ to become $\{1,3,2\}$, whereas the middle or second stratum has been "emptied" (and is deleted), and there remains only two strata. Strata are only emptied when the last remaining atomic stratum has been moved to another stratum.

To clarify how this works in the algorithm: each solution is represented by a vector of integers – atomic strata which have the same integer are in the same stratum. A separate vector of the unique integers in the solution represents the strata. For example, the first solution $\{1,3\},\{2\},\{4,5,6\}$ would be represented by the vector $[1\ \ 2\ \ 1\ \ 3\ \ 3\ \ 3]$ and the strata would be represented by the vector $[1\ \ 2\ \ 3]$. When the new solution is created, the second stratum has been removed and is no longer part of the solution. That is to say, the vector for the new solution is: $[1\ \ 1\ \ 1\ \ 3\ \ 3\ \ 3]$ and the strata vector is $[1\ \ 3]$. With stratum 2 removed, and for clarity, we rename stratum 3 to 2 so that this solution becomes: $[1\ \ 1\ \ 1\ \ 2\ \ 2\ \ 2]$, and the strata are now represented by the vector $[1\ \ 2]$. Strata $[1\ \ 2]$ will remain in any further solutions unless another stratum is "emptied" or a new stratum is added.

4.2 Evaluation and acceptance

Each new solution is evaluated using the Bethel-Chromy algorithm and the Metropolis acceptance criterion is applied. If accepted, the new solution differs from the previous solution only by the above mentioned perturbation. If it is not accepted, we continue with the previous solution, and again try moving q randomly chosen atomic strata between two randomly selected strata.

4.3 Sequences and new strata

This continues for the tunable length of the sequence, J. This should be long enough to allow the sequence to reach equilibrium. However, there is no rule to determine J. At the commencement of each new sequence, we have H strata in the current best solution. With a fixed probability of $1/J$, an additional stratum is added. If a new stratum is to be added, the SAA loops through each atomic stratum and moves it to a new stratum, which is called $H+1$, because each stratum is labelled sequentially from 1 to H (see Section 4.1), at a tunable probability, $P(H+1)$. The algorithm runs for a tunable number of sequences, *maxit*.

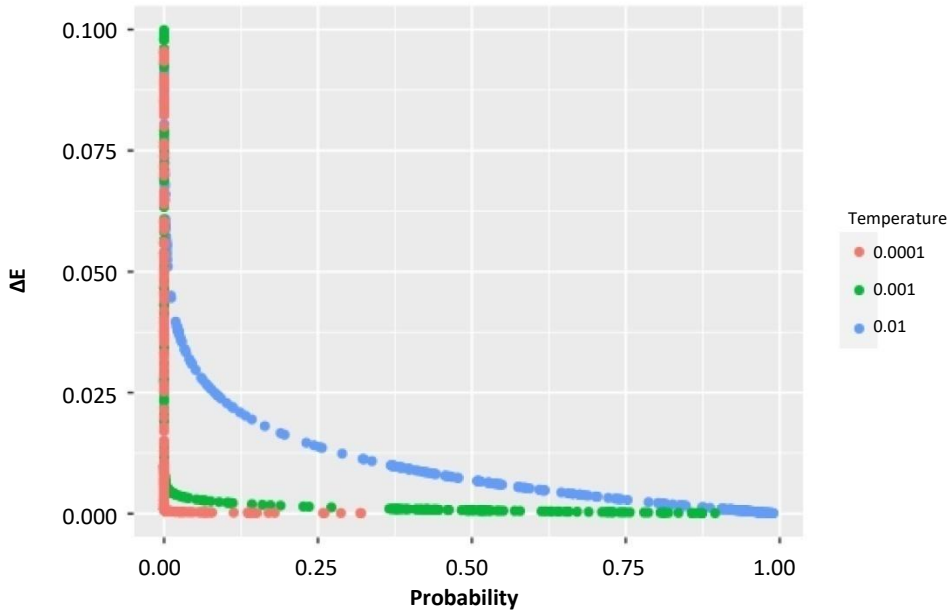
4.4 Temperature

The temperature is decremented from a starting temperature, $T_{\max}$, to a minimum temperature, $T_{\min}$, or until *maxit* has been reached. As we are starting with a near optimal solution, we select $T_{\max}$ as no greater than 0.01 and we set $T_{\min}$ to be 1.0×10^{-11}.

This is to allow for the advanced nature of the search, and allows the algorithm to focus more on the search for superior solutions, with an ever-reducing probability of accepting inferior solutions. However, a low temperature, T, does not always equate to a low probability of acceptance.

Small positive differences in solution quality (where the new solution has a marginally inferior quality to the current best solution), ΔE, occur often because we are starting with a good quality initial solution. Figure 4.1 demonstrates the probability of such solutions being accepted, $e^{\left(-\frac{\Delta E}{T}\right)}$, increases the smaller this difference becomes for the same T. Nonetheless, Figure 4.1 also demonstrates that for the same changes in solution quality as the T decreases, the probability also decreases (and it behaves increasingly like a hill climbing algorithm).

Figure 4.1 Probability of accepting an inferior solution as a function of ΔE and T.



5. Improving the performance of the simulated annealing algorithm using delta evaluation

As outlined earlier, the only difference between consecutive solutions is that q atomic strata have been moved from one group into another. As with the other heuristics, q is also tunable, and for the first sequence we have added the option of setting $q > 1$ and reducing q for each new solution in the first sequence until $q = 1$. The reason for this is that, where $q > 1$, the increased size of the perturbation can help reduce the number of strata. In this case, we set q as a tunable percentage of the solution size, or of the number of atomic strata, L, to be partitioned. After the first sequence $q = 1$.

Furthermore, as the strata are mutually exclusive, this movement of q atomic strata from one stratum to another does not affect the remaining strata in any way. Ross, Corne and Fang (1994) introduce a technique called delta evaluation, where the evaluation of a new solution makes use of previously evaluated similar solutions, to significantly speed up evolutionary algorithms/timetabling experiments. We use the similar properties of two consecutive solutions to apply delta evaluation to the SAA. It follows, therefore, that in the first sequence q should be kept low and the reduction to $q = 1$ should be swift.

The Bethel-Chromy algorithm requires the means and variances for each stratum in order to calculate the sample allocation. However, we use the information already calculated for the remaining $H - 2$ strata, and simply calculate for the two strata affected by the perturbation. Thus, the computation for the means and variances of the H strata is reduced to a mere subset of that otherwise required.

Now recall that the Bethel-Chromy algorithm starts with a default value for each α_g, and uses gradient descent to find a final value for each α_g. This search continues up to when the algorithm reaches a minimum step-size threshold, or alternatively exceeds a maximum number of iterations. This minimum

threshold is characterised by ϵ, which is set as 1.0×10^{-11} in Ballin and Barcaroli (2020), and the maximum number of iterations is 200. We make the assumption that this search will be substantially reduced if we use the α_g values from the evaluation of the current solution as a starting point for the next solution.

The above two implementations of delta evaluation result in a noticeable reduction in computation times as demonstrated in the experiments described below.

6. Comparing the performance of the two algorithms

6.1 Evaluation plan

In this section, we outline the comparison of the performance of the grouping genetic algorithm with the simulated annealing algorithm. We used a number of data sets of varying sizes in these experiments. There are a number of regions in each data set (labelled here as domains). An optimal stratification and minimum sample allocation was selected for each domain.

The sum of the samples for all domains provides the total sample size. The sample size, or cost of the solution, defines the solution quality. For more details on domains refer to Ballin and Barcaroli (2013). The aim of these experiments was to consider whether the SAA can attain comparable solution quality with the GGA in less computation time per solution thus resulting in savings in execution times.

However, we also compared the total execution times as this is a consequence of the need to train the hyperparameters for both algorithms. More details are available in the Appendix.

We tabulate the results of these experiments in Section 6.4 where for comparison purposes we express the SAA results as a ratio of those for the GGA.

6.2 Comparing the number of solutions generated

After the first iteration the GGA retains the elite solutions, E, from the previous iteration. These are calculated by the product of the elitism rate (the proportion of the chromosome population which are elite solutions), E_R, and the chromosome population size (the number of candidate solutions in each iteration), N_P. As E have already been evaluated they are not evaluated again.

For this reason, we compared the evaluation times for the evaluated solutions in the GGA with all those of the SAA. For the GGA, the total number of evaluated solutions, N_{GGAsol}, is a function of the number of domains, D, the chromosome population size, the non-elite solutions (calculated by the product of $1 - E_R$ and N_P), and the number of iterations, I. For more details on the implementation of GGAs (e.g. elite solutions, elitism rate, chromosome population) we refer the reader to (Falkenauer, 1998)

$$N_{\text{GGAsol}} = \left(D \times \left(N_P + \left(N_P \times \left(1 - E_R\right) \times \left(I - 1\right)\right)\right)\right).$$

For the simulated annealing algorithm, the maximum number of solutions, N_{SAAsol}, is the number of domains, D, by the number of sequences, *maxit*, by the length of sequence, J. Recall that the SAA also stops if the minimum temperature has been reached – hence we refer to the *maximum* number of solutions rather than the *total*. For comparability purposes however, because the temperature is decremented only at the end of each sequence and we have a small number of sequences in the experiments below we assume the full number of solutions has been generated

$$N_{\text{SAAsol}} = D \cdot maxit \cdot J.$$

6.3 Data sets, target and auxiliary variables

Table 6.1 provides a summary by data set of the target and auxiliary variables.

Table 6.1
Summary by data set of the target and auxiliary variables

Dataset	Target variables	Description	Auxiliary variables	Description
Swiss Municipalities	Surfacebois	wood area	POPTOT	total population
	Airbat	area with buildings	Hapoly	municipality area
American Community Survey, 2015	HINCP	Household income past 12 months	BLD	Units in structure
	VALP	Property value	TEN	Tenure
	SMOCP	Selected monthly owner costs	WKEXREL	Work experience of householder and spouse
	INSP	Fire/hazard/flood insurance yearly amount	WORKSTAT	Work status of householder or spouse in family households
			HFL	House heating fuel
			YBL	When structure first built
US Census, 2000	HHINCOME	total household income	PROPINSR	Annual property insurance cost
			COSTFUEL	annual home heating fuel cost
			COSTELEC	Annual electricity cost
			VALUEH	House value
Kiva Loans	term_in_months	duration for which the loan was disbursed	sector	high level categories, e.g. food
	lender_count	the total number of lenders	currency	currency of the loan
	loan	the amount in USD	activity	more granular category, e.g. fruits & vegetables
			region	region name within the country
			partner_id	ID of the partner organization
UN Commodity Trade Statistics data	trade_usd	value of the trade in USD	commodity	type of commodity e.g. “Horses, live except pure-bred breeding”
			flow	whether the commodity was an import, export, re-import or re-export
			category	category of commodity, e.g. silk or fertilisers

The target and auxiliary variables for the Swiss Municipalities data set were selected based on the experiment described in Ballin and Barcaroli (2020). Accordingly, *POPTOT* and *HApoly* were converted into categorical variables using the k-means clustering algorithm. However, we used more domains and iterations in our experiment. More information on this data set is provided by Barcaroli (2014).

For the remaining experiments we selected target and auxiliary variables which we deemed likely to be of interest to survey designers. Further details on the American Community Survey, 2015 (U.S. Census Bureau, 2016), the U.S. Census, 2000 (Ruggles, Genadek, Goeken, Grover and Sobek, 2017), Kiva Loans (Kiva, 2018), and the UN commodity trade statistics data (United Nations, 2017) metadata are available in O'Luing et al. (2019).

A further summary by data set of the number of records and atomic strata, along with a description of the domain variable, is provided in Table 6.2 below.

Table 6.2
Summary by data set of the number of records and atomic strata and a description of the domain variable

Data set	Number of records	Number of atomic strata, L	Domain variable
Swiss Municipalities	2,896	579	REG
American Community Survey, 2015	619,747	123,007	ST (the 51 states)
US Census, 2000	627,611	517,632	REGION
Kiva Loans	614,361	84,897	country code
UN Commodity Trade Statistics data	352,078	351,916	country or area

6.4 Results

As mentioned previously, we used an initial solution in each experiment that is created by the *KmeansSolution* algorithm (Ballin and Barcaroli, 2020). We then compared the performance of the algorithms in terms of average computation time (in seconds) per solution and solution quality. Table 6.3 provides the sample size, execution times and total execution times for the SAA and GGA.

Table 6.3
Summary by data set of the sample size and evaluation time for the grouping genetic algorithm and simulated annealing algorithm

Data set	GGA			SAA		
	Sample size	Execution time (seconds)	Total Execution time (seconds)	Sample size	Execution time (seconds)	Total Execution time (seconds)
Swiss Municipalities	128.69	753.82	10,434.30	125.17	248.91	8,808.63
American Community Survey, 2015	10,136.50	13,146.25	182,152.46	10,279.44	517.76	6,822.42
US Census, 2000	228.81	2,367.36	36,298.35	224.75	741.75	8,996.85
Kiva Loans	6,756.19	15,669.11	288,946.79	6,646.67	664.30	7,549.87
UN Commodity Trade Statistics data	3,216.68	6,535.97	88,459.22	3,120.07	1,169.26	12,161.80

The total execution time is the sum of the execution times for 20 evaluations of the GGA and SAA algorithms (by the MBO (model-based optimisation) function in the R package *mlrMBO* (Bischla, Richterb, Bossekc, Hornb, Thomasa and Langb, 2017)) using 20 sets of selected hyperparameters (i.e. one set for each evaluation). Details on the precision constraints and hyperparameters for each experiment can be found in the Appendix. Table 6.4 expresses the SAA results as a ratio of those for the GGA.

Table 6.4
Ratio comparison of the sample sizes, execution times, and total execution times for the grouping genetic algorithm and simulated annealing algorithm

Data set	Sample size	Execution time (seconds)	Total execution time (seconds)
Swiss Municipalities	0.97	0.33	0.84
American Community Survey, 2015	1.01	0.04	0.04
US Census, 2000	0.98	0.31	0.25
Kiva Loans	0.98	0.04	0.03
UN Commodity Trade Statistics data	0.97	0.18	0.14

As can be seen, the sample sizes are similar, however, the SAA shows significantly lower execution and total execution times. When these experiments are run in parallel, for cases where there is a large number of domains, there may not be enough cores to cover all domains in one run. Indeed, it may take several parallel runs to complete the task, and this will affect mean evaluation time. The computer specifications are provided in Table A.2. Table 6.5 shows the number of solutions evaluated by each algorithm to obtain the results shown in Table 6.3. It also provides a ratio comparison of the average execution time (in seconds) per solution.

Table 6.5
Number of solutions and ratio comparison of execution time (per second) between the grouping genetic algorithm and simulated annealing algorithm

Data set	Number of solutions evaluated		Average execution time per solution (seconds)		
	GGA	SAA	GGA	SAA	Proportion
Swiss Municipalities	840,140	210,000	0.0009	0.0012	1.3210
American Community Survey, 2015	2,550,510	459,000	0.0052	0.0011	0.2188
US Census, 2000	10,872	36,000	0.2177	0.0206	0.0946
Kiva Loans	2,190,730	730,000	0.0072	0.0009	0.1272
UN Commodity Trade Statistics data	2,395,026	1,539,000	0.0027	0.0008	0.2784

The above results indicate that the GGA has evaluated more solutions to find a solution of similar quality to the SAA in all cases, except for the *US Census, 2000* experiment. However, we also can see that the SAA takes less time to evaluate each solution in all cases except for the *Swiss Municipalities* experiment. The average execution time for each experiment can be considered in the context of the size

of the data set, parallelisation, and the particular sets of hyperparameters used for the GGA and SAA. In addition to this, there is also memoisation in the evaluation algorithm for the GGA, and the gains obtained by delta evaluation by the SAA.

Gains are more noticeable for larger data sets, because of the size of the solution and number of atomic strata in each stratum. As the strata get larger in size, the movement of q atomic strata from one stratum to another (where q is small) will have a smaller impact on solution quality and, therefore, the delta evaluation will be quicker.

7. Comparison with the continuous method in *SamplingStrata*

We also compared the SAA with the traditional genetic algorithm which Ballin and Barcaroli (2020) have applied to partition continuous strata. We used the target variables outlined in Table 6.1 above as both the continuous target and auxiliary variables (for clarity we outline them again in Table 7.1 below) along with the precision constraints outlined in Table A.1 (the Appendix). In practice, the target variable would not be exactly equal to the auxiliary variable though it is common for the auxiliary variable to be an imperfect version (for example an out-of-date or a related variable) available on the sampling frame. We invite the reader to consider this when reviewing the results of the comparisons below. It is also worth noting that initial solutions were created for both algorithms using the k-means method. Details on the training of hyperparameters for these experiments also can be found in the Appendix.

Table 7.1
Summary by data set of the target and auxiliary variable descriptions for the continuous method

Dataset	Target variables	Auxiliary variables	Description
Swiss Municipalities	Surfacebois	Surfacebois	wood area
	Airbat	Airbat	area with buildings
American Community Survey, 2015	HINCP	HINCP	Household income (past 12 months)
	VALP	VALP	Property value
	SMOCP	SMOCP	Selected monthly owner costs
	INSP	INSP	Fire/hazard/flood insurance (yearly amount)
US Census, 2000	HHINCOME	HHINCOME	total household income
Kiva Loans	term_in_months	term_in_months	duration for which the loan was disbursed
	lender_count	lender_count	the total number of lenders
	loan	loan	the amount in USD
UN Commodity Trade Statistics data	trade_usd	trade_usd	value of the trade in USD

The attained sample sizes are compared in Table 7.2 below where the sample size for the SAA is expressed as a ratio of the TGA. After the hyperparameters were fine-tuned (see Section A.6) the resulting sample sizes are comparable.

Table 7.2
Ratio comparison of the sample sizes for the traditional genetic algorithm and simulated annealing algorithm on the continuous method

Data set	TGA	SAA	Ratio
Swiss Municipalities	128.69	120.00	0.93
American Community Survey, 2015	4,197.68	3,915.48	0.93
US Census, 2000	192.71	179.89	0.93
Kiva Loans	3,062.33	3,017.79	0.99
UN Commodity Trade Statistics data	3,619.42	3,258.52	0.90

Table 7.3 compares the execution times for the set of hyperparameters that found the sample sizes for each algorithm in Table 7.2 above, as well as the total execution times taken to train that set of hyperparameters.

Table 7.3
Ratio comparison of the execution times and total execution times for the traditional genetic algorithm and simulated annealing algorithm on the continuous method

Data set	TGA		SAA		Ratio comparison	
	Execution time (seconds)	Total execution time (seconds)	Execution time (seconds)	Total execution time (seconds)	Execution time (seconds)	Total execution time (seconds)
Swiss Municipalities	753.82	10,434.30	213.44	1,905.82	0.28	0.18
American Community Survey, 2015	22,016.95	227,635.51	13,351.19	169,115.92	0.61	0.74
US Census, 2000	3,361.90	46,801.78	51.94	1,147.36	0.02	0.02
Kiva Loans	3,232.78	48,746.61	300.16	4,149.06	0.09	0.09
UN Commodity Trade Statistics data	29,045.23	326,931.63	73.18	1,287.38	0.003	0.004

These results indicate a significantly lower execution time for the SAA for the attained solution quality. The computational efficiency gained by delta evaluation in the training of the recommended hyperparameters is also evident in the total execution times. For the *American Community Survey, 2015* experiment significantly more solutions were generated by the SAA than the TGA as a result of the given hyperparameters and this impacts the execution and total execution times (see also Table 7.4). Table 7.4 compares the number of solutions generated by the traditional genetic algorithm with the simulated annealing algorithm.

Table 7.4
Comparison of the number of solutions generated by the traditional genetic algorithm and simulated annealing algorithm on the continuous method

Data set	Number of solutions evaluated	
	TGA	SAA
Swiss Municipalities	840,140	175,000
American Community Survey, 2015	918,102	5,100,000
US Census, 2000	43,272	18,000
Kiva Loans	146,730	292,000
UN Commodity Trade Statistics data	20,521,026	85,500

In all cases except for *Kiva Loans* and the *American Community Survey, 2015* the SAA has generated fewer solutions. The low number of solutions generated by both algorithms for the *US Census, 2000* experiment may indicate that the initial k-means solution was near the global minimum. The *American Community Survey, 2015* results indicate that the SAA generated significantly more solutions to get to a comparable sample size with the TGA. As we are moving, predominantly, $q = 1$ atomic strata between strata such changes in this case had limited impact on solution quality from one solution to the next. However, the gains achieved by delta evaluation meant that more solutions were evaluated per second leading to a more complete search and a lower sample size being attained.

For these experiments, the TGA took longer to find a comparable sample size in all cases. As pointed out in O'Luing et al. (2019), traditional genetic algorithms are not as efficient for grouping problems as the grouping genetic algorithm because solutions tend to have a great deal of redundancy. We would, therefore, propose that the GGA be applied also to continuous strata. On the basis of the above analysis, and the performance of SAAs in local search generally speaking along with the added gains in efficiency from delta evaluation, we would also propose that the SAA be considered as an alternative to the traditional genetic algorithm.

8. Conclusions

We compared the SAA with the GGA in the case of atomic strata and the TGA in the case of continuous strata (Ballin and Barcaroli, 2020). The k-means algorithm provided good starting points in all cases. When the hyperparameters have been fine-tuned all algorithms attain results of similar quality.

However, the execution times for the recommended hyperparameters are lower for the SAA than for the GGA with respect to atomic strata and traditional genetic algorithm with respect to continuous strata. Delta evaluation also has advantages in reducing the training times needed to find the suitable hyperparameters for the SAA.

The GGA might benefit from being extended into a memetic algorithm by using local search to quickly improve a chromosome before adding it to the GGA chromosome population.

The SAA, by using local search (along with a probabilistic acceptance of inferior solutions), is well suited to navigation out of local minima and the implementation of delta evaluation enables a more complete search of the local neighbourhood than would otherwise be possible in the same computation time.

9. Further work

The perturbation used by the SAA randomly moves q atomic strata, where mainly $q = 1$, from one stratum to another. This stochastic process is standard in default simulated annealing algorithms.

However, as we are using a starting solution where there is already similarity within the strata, this random process could easily move an atomic stratum $(q=1)$ to a stratum where it is less suited than the stratum it was in. This suggests the presence of a certain amount of redundancy in the search for the global minimum.

Lisic et al. (2018) conjecture that the introduction of nonuniform weighting in atomic strata selection could greatly improve performance of (their proposed) simulated annealing method by exchanging atomic strata near stratum boundaries more frequently than more important atomic strata. We agree that, for this algorithm, it would be more beneficial if there was a higher probability that an atomic stratum which was dissimilar to the other atomic strata was selected. We could then search for a more suitable stratum to move this atomic stratum to.

To achieve this we could first randomly select a stratum, and then measure the Euclidean distance of each atomic stratum from that stratum medoid, weighting the chance of selection of the atomic strata in accordance with their distance from the medoid. At this point, an atomic stratum is selected using these weighted probabilities.

The next step would be to use a K-nearest-neighbour algorithm to find the stratum medoid closest to that atomic stratum and move it to that stratum. This simple machine learning algorithm uses distance measures to classify objects based on their K nearest neighbours. In this case, $k=1$, so the algorithm in practice is a closest nearest neighbour classifier.

This additional degree of complexity to the algorithm may offset the gains achieved by using delta evaluation, particularly as the problem grows in size, thus reducing the number of solutions evaluated in the same running time. It might be more effective to use the column medians as an equivalent to the medoids. This could assist the algorithm find better quality solutions.

However, the above suggestions may only be effective at an advanced stage of the search, where the atomic strata in each stratum are already quite similar.

Acknowledgements

We wish to acknowledge the editorial staff and reviewers of *Survey Methodology* for their constructive suggestions in the review process for this journal submission, in particular the suggestion to compare the SAA with the traditional genetic algorithm in Ballin and Barcaroli (2020) in the case of continuous strata. This material is based upon work supported by the Insight Centre for Data Analytics and Science Foundation Ireland under Grant No. 12/RC/2289-P2 which is co-funded under the European Regional Development Fund. Also, this publication has emanated from research supported in part by a grant from Science Foundation Ireland under Grant number 16/RC/3918 which is co-funded under the European Regional Development Fund.

Appendix

Background details on the comparisons in Sections 6 and 7

A.1 Precision constraints

The target upper precision levels for these experiments, i.e. coefficients of variation, for each of the five experiments are provided in Table A.1 below.

Table A.1
Summary by data set of the upper limits for the coefficients of variation

Data set	CV
Swiss Municipalities	0.1
American Community Survey, 2015	0.05
US Census, 2000	0.05
Kiva Loans	0.05
UN Commodity Trade Statistics data	0.05

We selected an upper precision level of 0.1 for the *Swiss Municipalities* data set in keeping with the level set for the experiment in Ballin and Barcaroli (2020). We used an upper precision level of 0.05 for the remaining experiments, given that the upper CV levels generally set by national statistics institutes (NSIs) tend to be between 0.01 and 0.1, and, for this reason, results for CVs in the mid-point of this range are of interest.

A.2 Processing platform

Table A.2 below provides details of the processing platform used for these experiments.

Table A.2
Specifications of the processing platform

Specification	Details	Notes
Processor	AMD Ryzen 9 3950X 16-Core Processor, 3493 Mhz	
Cores	16 Core(s)	
Logical processors	32 Logical Processor(s)	32 cores in R
System model	X570 GAMING X	
System type	x64-based PC	
Installed physical memory (RAM)	16.0 GB	
Total virtual memory	35.7 GB	
OS name	Microsoft Windows 10 Pro	

In all cases, R version 4.0 or greater was used. We used the *foreach* (Microsoft Corporation and Weston, 2020a) and *doParallel* (Microsoft Corporation and Weston, 2020b) packages to run the experiments in parallel. The number of cores used in the experiments was 31 (32 less 1) and this means that in the three experiments with more than 31 domains (*American Community Survey 2015, Kiva Loans, UN Commodity Trade Statistics data*) the *foreach* algorithm continued to loop through the available cores until a solution had been found for all domains.

A.3 Hyperparameters for the grouping genetic algorithm and simulated annealing algorithm

Tables A.3 and A.4 below outline the number of domains in each experiment, along with number of iterations and chromosome population size for the grouping genetic algorithm and along with the number of sequences, length of sequence, and starting temperature for the simulated annealing algorithm. Section A.4 provides details on fine-tuning the hyperparameters. For more details on the hyperparameters of the GGA we refer the reader to Ballin and Barcaroli (2013) and O'Luing et al. (2019) and of the SAA to Sections 2.2 and 4.

Table A.3
Summary by data set of the hyperparameters for the grouping genetic algorithm for each domain

Data set	Domains	Number of iterations, I	Chromosome population size, N_p	Mutation chance	Elitism rate, E_R	Add strata factor
Swiss Municipalities	7	4,000	50	0.0053360	0.4	0.0037620
American Community Survey, 2015	51	5,000	20	0.0008134	0.5	0.0610529
US Census, 2000	9	100	20	0.0000007	0.4	0.0000472
Kiva Loans	73	3,000	20	0.0007221	0.5	0.0685005
UN Commodity Trade Statistics data	171	1,000	20	0.0004493	0.3	0.0866266

Table A.4
Summary by data set of the hyperparameters for the simulated annealing algorithm for each domain

Data set	Domains	Number of sequences, *maxit*	Length of sequence, *J*	Temperature, *T*	Decrement constant, DC	% of L for maximum q value, $L_{\max\%}$	Probability of new stratum, $P(H+1)$
Swiss Municipalities	7	10	3,000	0.0000720	0.5083686	0.0183356	0.0997907
American Community Survey, 2015	51	3	3,000	0.0002347	0.6873029	0.0076477	0.0291729
US Census, 2000	9	2	2,000	0.0006706	0.5457192	0.0189395	0.0806919
Kiva Loans	73	5	2,000	0.0009935	0.7806557	0.0143925	0.0317491
UN Commodity Trade Statistics data	171	3	3,000	0.0007902	0.5072737	0.0234728	0.0013775

A.4 Fine-tuning the hyperparameters for the grouping genetic algorithm and simulated annealing algorithm

In order to fine-tune the initial parameters or *hyperparameters* we used sequential model-based optimization (Hutter, Hoos and Leyton-Brown, 2010). We first generated an initial design of hyperparameters from the value ranges described for the GGA in Table A.5 and in Table A.6 for the SAA below using the latin hypercube design method (McKay, Beckman and Conover, 2000).

Table A.5
Ranges for fine-tuning the hyperparameters for the grouping genetic algorithm

	Iterations			Population size			Mutation chance		Elitism rate, E_R			Add strata factor	
Value type	**Discrete**			**Discrete**			**Numeric**		**Discrete**			**Numeric**	
Value range	**Lower value**	**Upper value**	**Increments**	**Lower value**	**Upper value**	**Increments**	**Lower value**	**Upper value**	**Lower value**	**Upper value**	**Increments**	**Lower value**	**Upper value**
Swiss Municipalities	500	5,000	500	10	50	10	0	0.10	0.1	0.5	0.1	0	0.1
American Community Survey, 2015	1,000	5,000	1,000	10	20	10	0	0.001	0.1	0.5	0.1	0	0.1
Kiva Loans	1,000	3,000	1,000	10	20	10	0	0.001	0.1	0.5	0.1	0	0.1
UN Commodity Trade Statistics data	500	1,000	500	10	20	10	0	0.001	0.1	0.5	0.1	0	0.1
US Census, 2000	50	100	50	10	20	10	0	0.000001	0.1	0.5	0.1	0	0.0001

Table A.6
Ranges for fine-tuning the hyperparameters for the simulated annealing algorithm

	Number of sequences, *maxit*			Length of sequence, J			Temperature, T		Decrement constant, DC		% L for maximum q value, $L_{max\%}$		Probability of new stratum, $P(H+1)$	
Value type	**Discrete**			**Discrete**			**Numeric**		**Numeric**		**Numeric**		**Numeric**	
Value range	**Lower value**	**Upper value**	**Increments**	**Lower value**	**Upper value**	**Increments**	**Lower value**	**Upper value**	**Lower value**	**Upper value**	**Lower value**	**Upper value**	**Lower value**	**Upper value**
Swiss Municipalities	10	50	10	1,000	3,000	1,000	0	0.001	0.5	1	0.0001	0.025	0	0.1
American Community Survey, 2015	1	3	1	1,000	3,000	1,000	0	0.001	0.5	1	0.0001	0.025	0	0.1
Kiva Loans	1	5	1	1,000	2,000	1,000	0	0.001	0.5	1	0.0001	0.025	0	0.1
UN Commodity Trade Statistics data	1	3	1	1,000	3,000	1,000	0	0.001	0.5	1	0.0001	0.025	0	0.1
US Census, 2000	1	2	1	1,000	2,000	1,000	0	0.001	0.5	1	0.0001	0.025	0	0.1

As some of the hyperparameter value ranges were discrete, we used a random forest with regression trees to develop a surrogate learner model. After this, a confidence bound using a lambda value, λ, to control the trade-off between exploitation and exploration was used as the acquisition function. The focus

search approach (Bischla et al., 2017) was used to optimise the acquisition function which, in turn, was used to propose the hyperparameters which were evaluated using the surrogate function (which is a cheaper alternative to using the GGA or SAA algorithms). From these, the most promising hyperparameters were then evaluated by the GGA or SAA and the hyperparameters and solution costs added to the initial design. The process was then repeated for a set number of iterations and the best performing hyperparameters and solution outcomes were selected. We implemented this using the *MBO* function with the parameters outlined in Table A.7. These are distinct from the parameters being fine-tuned, which are outlined in Tables A.5 and A.6 above.

Table A.7
Parameters used in the MBO Function

MBO parameters	Value
Initial Design size (Latin Hypercube Design method)	10
Iterations, number of	10
Number of Trees	500
Lambda, λ	5
Focus Search Points	1,000

As can be seen from the limited scope of the *MBO* function parameters this was not an exhaustive fine-tuning of the hyperparameters for the GGA and SAA. The aim of these experiments was to consider whether the SAA can attain comparable solution quality with the GGA in less computation time per solution thus resulting in savings in execution times. However, we also compared the total execution times as this is a consequence of the need to train the hyperparameters for both algorithms.

Tables outlining the hyperparameters, in each of the 20 fine-tuning iterations, for each experiment are available from the authors on request. The first 10 sets of hyperparameters were randomly generated from the ranges laid out in Tables A.5 and A.6. The ranges selected were identified using practical knowledge of the algorithms and data. The second 10 sets reflects the *MBO* function's attempts to learn the hyperparameters that best lead each algorithm towards the optimal solution using the previous solutions as a guide.

A.5 Hyperparameters for the traditional genetic algorithm and simulated annealing algorithm

Tables A.8 and A.9 outline the hyperparameters for the tradtional genetic algorithm and the simulated annealing algorithm. The add strata factor option is not available for the traditional genetic algorithm and, therefore, is not included in Table A.8. More details on fine-tuning the hyperparameters are provided in Section A.6.

Table A.8
Hyperparameters for the traditional genetic algorithm

Data set	Iterations	Population size	Mutation chance	Elitism rate, E_R
Swiss Municipalities	4,000	50	0.0053360	0.4
American Community Survey, 2015	1,000	20	0.0009952	0.1
US Census, 2000	400	20	0.0002317	0.4
Kiva Loans	200	20	0.0817285	0.5
UN Commodity Trade Statistics data	5,000	30	0.0005599	0.2

Table A.9
Hyperparameters for the simulated annealing algorithm

Data set	Number of sequences, *maxit*	Length of sequence, J	Temperature, T	Decrement constant, DC	% for maximum q value, $L_{\max\%}$	Probability of new stratum, $P(H+1)$
Swiss Municipalities	5	5,000	0.02311057	0.9427609	0.3736443	0.0229361
American Community Survey, 2015	50	2,000	0.00000005	0.9528952	0.0001021	0.0000008
US Census, 2000	1	2,000	0.00002000	0.9665631	0.0221147	0.0160408
Kiva Loans	2	2,000	0.00053839	0.8660943	0.0014281	0.0216320
UN Commodity Trade Statistics data	2	250	0.00067481	0.9309940	0.0203113	0.0149499

A.6 Fine-tuning the hyperparameters for the traditional genetic algorithm and simulated annealing algorithm

We fine-tuned the hyperparameters for the TGA and SAA using the same methodology described in Section A.4. Tables outlining the hyperparameters, in each of the 20 fine-tuning iterations, for each experiment are available from the authors on request. The first 10 sets were randomly generated using practical knowledge of the algorithms and data to define upper and lower bounds for each hyperparameter. In the second 10 sets the MBO function attempts to optimise the hyperparameters using the previous solutions as a guide.

References

Ballin, M., and Barcaroli, G. (2013). Joint determination of optimal stratification and sample allocation using genetic algorithm. *Survey Methodology*, 39, 2, 369-393. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2013002/article/11884-eng.pdf.

Ballin, M., and Barcaroli, G. (2020). Optimization of sampling strata with the SamplingStrata package. Accessed 3 May 2021. https://barcaroli.github.io/SamplingStrata/articles/SamplingStrata.html.

Barcaroli, G. (2014). SamplingStrata: An R package for the optimization of stratified sampling. *Journal of Statistical Software*, 61, 4, 1-24.

Benedetti, R., Espa, G. and Lafratta, G. (2008). A tree-based approach to forming strata in multipurpose business surveys. *Survey Methodology*, 34, 2, 195-203. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2008002/article/10760-eng.pdf.

Bethel, J.W. (1985). An optimum allocation algorithm for multivariate surveys. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 209-212.

Bethel, J. (1989). Sample allocation in multivariate surveys. *Survey Methodology*, 15, 1, 47-57. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1989001/article/14578-eng.pdf.

Bischla, B., Richterb, J., Bossekc, J., Hornb, D., Thomasa, J. and Langb, M. (2017). mlrmbo: A modular framework for model-based optimization of expensive black-box functions. *Gradient Boosting in Automatic Machine Learning: Feature Selection and Hyperparameter Optimization*, 36.

Černỳ, V. (1985). Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *Journal of Optimization Theory and Applications*, 45, 1, 41-51.

Chromy, J.R. (1987). Design optimization with multiple objectives. *Proceedings of the Survey Research Methods Section*.

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

Cortez, P. (2014). *Modern Optimization with R*. Springer.

Falkenauer, E. (1998). *Genetic Algorithms and Grouping Problems*. New York: John Wiley & Sons, Inc.

Grover, L.K. (1987). Standard cell placement using simulated sintering. In *Proceedings of the 24th ACM/IEEE Design Automation Conference*, 56-59.

Hartigan, J.A., and Wong, M.A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 28, 1, 100-108.

Hoos, H.H., and Stützle, T. (2004). *Stochastic Local Search: Foundations and Applications*. Elsevier.

Hutter, F., Hoos, H.H. and Leyton-Brown, K. (2010). Sequential model-based optimization for general algorithm configuration (extended version). Technical Report TR-2010–10, University of British Columbia, Computer Science.

Johnson, D.S., Aragon, C.R., McGeoch, L.A. and Schevon, C. (1989). Optimization by simulated annealing: An experimental evaluation; part I, graph partitioning. *Operations Research*, 37, 6, 865-892.

Kirkpatrick, S., Gelatt, C.D. and Vecchi, M.P. (1983). Optimization by simulated annealing. *Science*, 220, 4598, 671-680.

Kiva (2018, Mar). *Data Science for Good: Kiva Crowdfunding*. https://www.kaggle.com/kiva/data-science-for-good-kiva-crowdfunding.

Lisic, J., Sang, H., Zhu, Z. and Zimmer, S. (2018). Optimal stratification and allocation for the june agricultural survey. *Journal of Official Statistics*, 34, 1, 121-148.

Luke, S. (2013). *Essentials of Metaheuristics* (2 Ed.). Lulu. Available for free at http://cs.gmu.edu/~sean/book/metaheuristics/.

McKay, M.D., Beckman, R.J. and Conover, W.J. (2000). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 42, 1, 55-61.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21, 6, 1087-1092.

Microsoft Corporation and Weston, S. (2020a). *foreach: Provides Foreach Looping Construct*. R package version 1.5.1.

Microsoft Corporation and Weston, S. (2020b). *doParallel: Foreach Parallel Adaptor for the 'Parallel' Package*. R package version 1.0.16.

O'Luing, M., Prestwich, S. and Tarim, S.A. (2019). A grouping genetic algorithm for joint stratification and sample allocation designs. *Survey Methodology*, 45, 3, 513-531. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2019003/article/00007-eng.pdf.

Rose, J.S., Snelgrove, W.M. and Vranesic, Z.G. (1988). Parallel standard cell placement algorithms with quality equivalent to simulated annealing. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 7, 3, 387-396.

Ross, P., Corne, D. and Fang, H.-L. (1994). Improving evolutionary timetabling with delta evaluation and directed mutation. In *International Conference on Parallel Problem Solving from Nature*, 556-565. Springer.

Ruggles, S., Genadek, K., Goeken, R., Grover, J. and Sobek, M. (2017). Integrated public use microdata series: Version 7.0 [dataset]. Minneapolis: University of Minnesota.

Sörensen, K., and Glover, F. (2013). Metaheuristics. *Encyclopedia of Operations Research and Management Science*, 62, 960-970.

Takeang, C., and Aurasopon, A. (2019). Multiple of hybrid lambda iteration and simulated annealing algorithm to solve economic dispatch problem with ramp rate limit and prohibited operating zones. *Journal of Electrical Engineering & Technology*, 14, 1, 111-120.

United Nations (2017, Nov). *Global Commodity Trade Statistics*. https://www.kaggle.com/unitednations/global-commodity-trade-statistics.

U.S. Census Bureau (2016). *2015 ACS Public Use Microdata Sample (PUMS)*. Washington, D.C. https://factfinder.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t#.

Van Laarhoven, P.J., and Aarts, E.H. (1987). Simulated annealing. In *Simulated Annealing: Theory and Applications*, Springer.

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 37, No. 4, December 2021

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 38, No. 1, March 2022
Special Issue on Price Indices in Official Statistics

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 49, No. 3, September/septembre 2021

The Canadian Journal of Statistics **La revue canadienne de statistique**

CONTENTS TABLE DES MATIÈRES

Volume 49, No. 4, December/décembre 2021

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* as a guide and note particularly the points below. Articles must be submitted in Word or Latex, preferably in Word with MathType for the mathematical expressions. A pdf version is also 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 and Introduction

2.1 The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.
2.2 The last paragraph of the introduction should contain a brief description of each section.

3. Style

3.1 Avoid footnotes and abbreviations.
3.2 Limit the use of acronyms. If an acronym is used, it must be defined the first time it occurs in the paper.
3.3 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as "exp(·)" and "log(·)", etc.
3.4 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.
3.5 Bold fonts should normally be used to distinguish vectors and matrices from scalars.

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 top of tables or figures. 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 first time a reference is cited in the text, the name of all authors must be written. For subsequent occurrences, the names of all authors can again be written. However, if the reference contains three or more authors, the names of the second and subsequent authors can be replaced with "et al.".
5.3 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, including tables, figures and references.