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

Survey Methodology

Survey Methodology
47-1

Release date: June 24, 2021

Statistics Canada Statistique Canada

Canada

How to obtain more information

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

You can also contact us by

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

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

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

Depository Services Program

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

Standards of service to the public

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

Note of appreciation

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

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 2021

•

Volume 47

•

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@canada.ca).

Survey Methodology

A Journal Published by Statistics Canada

Volume 47, Number 1, June 2021

Contents

Survey Methodology, June 2021
Vol. 47, No. 1, pp. 1-2
Statistics Canada, Catalogue No. 12-001-X

Waksberg Invited Paper Series

The journal *Survey Methodology* has established in 2001 an annual invited paper series in honor of the late Joseph Waksberg to recognize his outstanding contributions to survey statistics and methodology. Each year a prominent survey statistician is chosen by a four-person selection committee appointed by *Survey Methodology* and the *American Statistical Association*. The selected statistician is invited to write a paper for *Survey Methodology* that reviews the development and current state of an important topic in the field of survey statistics and methodology. The paper reflects the mixture of theory and practice that characterized Joseph Waksberg's work.

This issue of *Survey Methodology* opens with the 20th paper of the Waksberg Invited Paper Series. The editorial board would like to thank the members of the selection committee Michael A. Hidiroglou (Chair), Bob Fay, Jean Opsomer and Elizabeth Stuart for having selected Roger Tourangeau as the author of 2020 Waksberg Award paper.

2020 Waksberg Invited Paper

Author: Roger Tourangeau

Roger Tourangeau is a former Vice President at Westat, where he co-directed the Methodology Unit. Before joining Westat, he was the Director of the Joint Program in Survey Methodology at the University of Maryland and a Research Professor at the University of Michigan's Institute for Social Research. Tourangeau is the lead author on two books—*The Psychology of Survey Response* (with Lance Rips and Kenneth Rasinski), the winner of the 2006 American Association for Public Opinion Research (AAPOR) Book Award, and *The Science of Web Surveys* (with Frederick Conrad and Mick Couper). He has published more than 90 articles on survey methods topics. His work has been supported by grants from the National Science Foundation, the National Institute on Drug Abuse, and the National Institute of Child Health and Human Development. He has won numerous awards and honors. He was made a Fellow of the American Statistical Association (ASA) in 1999; received the Helen Dinerman Award (the highest honor given by the World Association for Public Opinion Research) in 2002; received the AAPOR Innovators Award (with Thomas Jabine, MironStraf, and Judy Tanur) in 2006;chaired the Survey Research Methods Section of the ASA in 2006; co-founded the *Journal of Survey Statistics and Methodology* in 2012; served as AAPOR's President in 2016-2017; was the 2018 Morris Hansen Lecturer (sponsored by the Washington Statistical Society); and was named the 2020 Sirken Lecturer by AAPOR and ASA.

Waksberg Award honorees and their invited papers since 2001

2021 Sharon **Lohr**, Manuscript topic under consideration, (expected for vol. 47, 2).

2020 Roger **Tourangeau**, "Science and survey management". *Survey Methodology* , vol. 47, 1, 3-28.

2019 Chris **Skinner**.

2018 Jean-Claude **Deville**, "De la pratique à la théorie : l'exemple du calage à poids bornés". 10ème Colloque francophone sur les sondages, Université Lumière Lyon 2.

2017 Donald **Rubin**, "Conditional calibration and the sage statistician". *Survey Methodology*, vol. 45, 2, 187-198.

2016 Don **Dillman**, "The promise and challenge of pushing respondents to the Web in mixed-mode surveys". *Survey Methodology*, vol. 43, 1, 3-30.

2015 Robert **Groves**, "Towards a quality framework for blends of designed and organic data". Proceedings: *Symposium 2016, Growth in Statistical Information: Challenges and Benefits*.

2014 Constance **Citro**, "From multiple modes for surveys to multiple data sources for estimates". *Survey Methodology*, vol. 40, 2, 137-161.

2013 Ken **Brewer**, "Three controversies in the history of survey sampling". *Survey Methodology*, vol. 39, 2, 249-262.

2012 Lars **Lyberg**, "Survey quality". *Survey Methodology*, vol. 38, 2, 107-130.

2011 Danny **Pfeffermann**, "Modelling of complex survey data: Why model? Why is it a problem? How can we approach it?". *Survey Methodology*, vol. 37, 2, 115-136.

2010 Ivan **Fellegi**, "The organisation of statistical methodology and methodological research in national statistical offices". *Survey Methodology*, vol. 36, 2, 123-130.

2009 Graham **Kalton**, "Methods for oversampling rare subpopulations in social surveys". *Survey Methodology*, vol. 35, 2, 125-141.

2008 Mary **Thompson**, "International surveys: Motives and methodologies". *Survey Methodology*, vol. 34, 2, 131-141.

2007 Carl-Erik **Särndal**, "The calibration approach in survey theory and practice". *Survey Methodology*, vol. 33, 2, 99-119.

2006 Alastair **Scott**, "Population-based case control studies". *Survey Methodology*, vol. 32, 2, 123-132.

2005 J.N.K. **Rao**, "Interplay between sample survey theory and practice: An appraisal". *Survey Methodology*, vol. 31, 2, 117-138.

2004 Norman **Bradburn**, "Understanding the question-answer process". *Survey Methodology*, vol. 30, 1, 5-15.

2003 David **Holt**, "Methodological issues in the development and use of statistical indicators for international comparisons". *Survey Methodology*, vol. 29, 1, 5-17.

2002 Wayne **Fuller**, "Regression estimation for survey samples". *Survey Methodology*, vol. 28, 1, 5-23.

2001 Gad **Nathan**, "Telesurvey methodologies for household surveys – A review and some thoughts for the future". *Survey Methodology*, vol. 27, 1, 7-31.

Survey Methodology, June 2021
Vol. 47, No. 1, pp. 3-28
Statistics Canada, Catalogue No. 12-001-X

Science and survey management

Roger Tourangeau[1]

Abstract

It is now possible to manage surveys using statistical models and other tools that can be applied in real time. This paper focuses on three developments that reflect the attempt to take a more scientific approach to the management of survey field work: 1) the use of responsive and adaptive designs to reduce nonresponse bias, other sources of error, or costs; 2) optimal routing of interviewer travel to reduce costs; and 3) rapid feedback to interviewers to reduce measurement error. The article begins by reviewing experiments and simulation studies examining the effectiveness of responsive and adaptive designs. These studies suggest that these designs can produce modest gains in the representativeness of survey samples or modest cost savings, but can also backfire. The next section of the paper examines efforts to provide interviewers with a recommended route for their next trip to the field. The aim is to bring interviewers' field work into closer alignment with research priorities while reducing travel time. However, a study testing this strategy found that interviewers often ignore such instructions. Then, the paper describes attempts to give rapid feedback to interviewers, based on automated recordings of their interviews. Interviewers often read questions in ways that affect respondents' answers; correcting these problems quickly yielded marked improvements in data quality. All of the methods are efforts to replace the judgment of interviewers, field supervisors, and survey managers with statistical models and scientific findings.

Key Words: Survey management; Responsive design; Adaptive design; Optimal routing.

1. Introduction

Surveys are in trouble these days, faced with the twin dilemmas of rising costs and falling response rates (e.g., Tourangeau, 2017; Williams and Brick, 2018). Both trends have been apparent in the United States since the 1970s (Atrostic, Bates, Burt and Silberstein, 2001; Steeh, Kirgis, Cannon and Dewitt, 2001), but seem to have accelerated in the last ten years or so. The same trends hold throughout the developed world (de Leeuw and de Heer, 2002). It seems fair to say that survey researchers do not really know what hit them (although see Brick and Williams (2013), for a thoughtful exploration of the possible causes behind these trends). But it is clear that fewer and fewer people want to do surveys these days; the downward trend in response rates mainly reflects increasing resistance to surveys among members of the general public.

Partly in response to this global industry-wide crisis, researchers have taken a closer look at the impact of falling response rates on the accuracy of survey estimates and have also proposed various measures to counter declining response rates. For example, more and more surveys have begun to offer incentives, make use of advance letters, and increase the number of contact attempts they make.

But another trend has been the use of a range of methods to improve the *management* of surveys to reduce the potential for error, data collection costs, or both. In Section 2, we review these efforts, generally known as *responsive* and *adaptive* designs. In Section 3, we look at another method for reducing cost and increasing efficiency in face-to-face surveys. This method – *optimal routing* – involves survey managers giving field interviewers detailed instructions about which cases to try to interview and what

1. Roger Tourangeau, 1601 Third Avenue, Apt. 19AW, New York, NY 10128. E-mail: RTourang@gmail.com.

route to follow in their next venture into the field. In Section 4, we look at another development with the potential to improve the performance of interviewers with the computer audio-recording of interviews, or *CARI* (Hicks, Edwards, Tourangeau, McBride, Harris-Kojetin and Moss, 2010). CARI allows central office staff the opportunity to hear how the interviewers are administering the questions in the field and make midcourse corrections in their performance. Research has shown that field interviewers depart from script more often than telephone interviewers do (Schaeffer, Dykema and Maynard, 2010; West and Blom, 2017), presumably because telephone interviewers can be monitored and given feedback in real time. In this fourth section, we describe two experiments in which central office staff provided *rapid feedback* to field interviewers – feedback provided within two or three days of the interview. What these techniques have in common is replacing the judgment of interviewers and field staff with the evidence- based prescriptions of survey managers – that is, they are attempts to replace management art with management science. Finally, Section 5 presents some conclusions.

2. Responsive and adaptive design

Responsive and adaptive designs refer to a family of methods for tailoring field work to reduce bias, variance, or cost (see Chun, Heeringa and Schouten (2018); Schouten, Peytchev and Wagner (2017); and Tourangeau, Brick, Lohr and Li (2017), for reviews). With responsive designs, researchers use multiple phases of data collection to reduce survey costs or errors. Adaptive designs use various forms of case prioritization, tailoring, and rules for stopping data collection to achieve similar goals.

Groves and Heeringa (2006) got this particular ball rolling with their description of responsive designs:

> Responsive designs are organized about design phases. A design phase is a time period of a data collection during which the same set of sampling frame, mode of data collection, sample design, recruitment protocols, and measurement conditions are extant. For example, a survey may start with a mail questionnaire attempt in the first phase, follow it with a telephone interview phase on non-respondents to the first phase and then have a final third phase of face-to-face interviewing. … Note that this use of "phase" includes more design features than merely the sample design, which are common to the term "multi-phase sampling". (Pages 440-441)

Of course, the American Community Survey had been using a three-phase design (mail followed by telephone follow-up followed by face-to-face follow-up with a subsample of the remaining cases) just like the one Groves and Heeringa described years before Groves and Heeringa dubbed these "responsive designs" (U.S. Census Bureau, 2014).

Groves and Heeringa cite several surveys that used responsive designs but focus mainly on Cycle 6 of the National Survey of Family Growth (NSFG). Most of the surveys they discuss, including Cycle 6 of the NSFG, applied two-phase sampling (that is, they selected a subsample of the nonrespondents remaining at a certain point in the field period and restricted further follow-up to this subsample) and offered larger incentives or made other changes to the data collection protocol for these final-phase cases. The real

innovation in the NSFG was not in its use of multiple phases of sampling (which had been around since Hansen and Hurwitz (1946)) or multiple modes of data collection (in fact, in the NSFG, all the cases were interviewed face-to-face) but in the application of paradata and real-time propensity modeling to guide the field work. The subsampling of nonrespondents in the Cycle 6 of the NSFG was based on propensity models that were updated frequently and that incorporated information gleaned from prior contacts with the sample case. In the final phase of Cycle 6 the NSFG, data collection was restricted to certain sample areas, with areas with larger numbers of active cases and those with cases with relatively high estimated propensities more likely to be retained for further follow-up field work.

Another difference between responsive designs and more traditional multi-phase designs, at least conceptually, is the notion of phase capacity. Groves and Heeringa argue that a given phase of data collection approaches a limit in its ability to change the survey estimates (and reduce any biases). Once it reaches this capacity limit, a change in protocol may be needed to improve the representativeness of the sample and reduce bias. Ideally, the later phases of data collection bring in different types of respondents from the earlier phases, reducing any remaining nonresponse biases. Different types of people may be inclined to respond by mail from those who respond to a face-to-face interview; larger incentives may help recruit those who are not interested in the topic (Groves, Singer and Corning, 2000). In the best case, the different phases of data collection are complementary and, together, create a more representative sample than each of the individual phases.

2.1 Case prioritization and related strategies

Cycle 6 of the NSFG is an early example of a strategy known as *case prioritization* – deliberately allocating more effort to some sample cases than to others. Of course, survey managers have always given priority to some cases over others. Interviewers are instructed to make sure they keep appointments, for example, or to set "soft" refusal cases aside for a while. What is different about the recent uses of case prioritization is that they are not based on a case's disposition but on models of the case's response propensity. In the Cycle 6 of the NSFG, a probability subsample of cases was kept for further work, with the second phase sampling probabilities partly based on the predicted propensities of the remaining cases. Later efforts have been explicit in their use of response propensities to guide the field work.

Depending on which cases are prioritized, case prioritization can serve a variety of goals. For example, focusing field work on cases with high response propensities may maximize the final sample size or reduce the costs per case. Beaumont, Bocci and Haziza (2014) distinguish three potential goals for such designs:

1) Minimizing variance;
2) Minimizing nonresponse bias or some proxy for it, such as sample imbalance (Särndal, 2011; see also Schouten, Cobben and Bethlehem, 2009); or
3) Maximizing response rates.

The first and third goals are related in that maximizing response rates tends to produce larger samples and, as a result, lower sample variances. Although some researchers have begun looking at the use of such designs to reduce measurement errors (Calinescu, Bhulai and Schouten, 2013), most efforts to date have been attempts to reduce nonresponse bias or costs.

With Cycle 6 of the NSFG, it is not completely clear what the statistical goal was. Oversampling areas with larger numbers of remaining cases and those with higher-propensity cases would tend to maximize the final sample size and reduce costs per case. Consistent with this, Groves, Benson, Mosher, Rosenbaum, Granda, Axinn, Lepkowski and Chandra (2005) noted that "this design option placed large emphasis on the cost efficiency of the … [final] phase design to produce interviews, not on minimizing standard errors of the resulting data set". However, Groves et al. (2005) also said that the final phase of data collection was intended to produce a "more representative" sample (page 38) by altering the data collection protocol to appeal to sample members who had failed to respond earlier. However, targeting areas with more cases with high estimated response propensities – that is, the cases predicted to be easiest to get – might actually exacerbate any problems with representativeness by bringing in additional respondents similar to those who had already responded.

Most later applications of case prioritization have taken the opposite tack, attempting to equalize the overall response propensities by focusing the field effort on the hardest cases. To see why this is a reasonable strategy, it is useful to take a closer look at the mathematics of nonresponse bias.

2.2 Factors affecting nonresponse bias

Under a stochastic perspective (e.g., Bethlehem, 1988), the bias of the unadjusted estimator of a mean or proportion $(\hat{\bar{y}})$ can be expressed as

$$\text{Bias}(\hat{\bar{y}}) \approx \frac{\sigma_\phi \sigma_y \rho_{\phi,y}}{\bar{\phi}}, \tag{2.1}$$

where $\bar{\phi}$ and σ_ϕ are the mean and standard deviation of the response propensities, σ_y is the standard deviation of a survey variable, and $\rho_{\phi,y}$ is the correlation between the response propensities and that survey variable. As (2.1) clearly demonstrates, both the overall response rate $(\bar{\phi})$ and the variation in the response rates (σ_ϕ) play a role in the bias, so that trying to maximize the response rates (e.g., by prioritizing the relatively easy cases) or to equalize the response propensities (by prioritizing the harder cases) are both reasonable things to do.

As a number of researchers have pointed out, nonresponse bias is a property of a survey estimate not of a survey, and, as (2.1) makes explicit, two variable-level properties also affect the bias – the correlation between the survey variable and the response propensities $(\rho_{\phi,y})$ and the variability of the survey variable (σ_y), both of which vary from one survey variable to the next. Given that two of the ingredients in the bias expression are study-level factors and two are variable-level, the question arises how much of the variation in nonresponse bias is between surveys and how much is within surveys.

Brick and I (Brick and Tourangeau, 2017) attempted to address this issue by reanalyzing data from a study done by Groves and Peytcheva (2008). They examined 959 nonresponse bias estimates from 59 studies. Eight hundred and four of these bias estimates involved proportions; almost all the others were means. (Four of the estimates seemed problematic to us, so we dropped them from our reanalysis.) Like Groves and Peytcheva, we examined the absolute relative bias statistic (absolute relbias), or the absolute

difference between the respondent estimate and the full sample estimate divided by the full sample estimate:

$$R_i = \frac{|\theta_{ri} - \theta_{ni}|}{\theta_{ni}}, \quad (2.2)$$

in which R_i is the absolute relbias for statistic i, θ_{ri} is the estimated value for that statistic based on the respondents, and θ_{ni} is the corresponding full sample estimate. The absolute relbias is useful in that it puts all the bias estimates on the same metric the percentage by which the estimate is off. Our reanalysis also examined the absolute differences (the numerator in (2.2)) for the estimated proportions.

Table 2.1 displays various statistics from the reanalysis. For example, we calculated the correlation between the individual bias estimates and the study-level response rates; these results are shown in the top panel of the table. The middle three panels of the table show what happens when the average bias from the study is used in place of the individual bias estimates. Some of the correlations based on study-level averages are considerably higher than those based on the individual estimates, particularly when the data are weighted by the number of estimates from each study (r's of 0.40 to 0.55). The bottom two panels of the table show that there is a substantial study-level component to the nonresponse bias. For example, the R^2 estimates from a one-way ANOVA indicate that the between-study component accounts for 21 to 40 percent of the overall variation in the nonresponse bias estimates. The results from multi-level models lead to similar conclusions. This between-study component of the bias presumably reflects two main variables – the mean response propensity (reflected in the overall response rate) and the variation across respondents in the response propensities.

Table 2.1
Relationship between response rates and bias measures at the estimate and study level

	All statistics	Proportions only
Estimate-level correlations		
Response rate and absolute relbias	-0.191(n = 955)	-0.256 (n = 802)
Response rate and absolute difference	-	-0.323 (n = 802)
Unweighted study-level correlations		
Response rate and mean absolute relbias	-0.255 (n = 57)	-0.315 (n = 43)
Response rate and mean absolute difference	-	-0.246 (n = 43)
Study-level correlations weighted by number of estimates		
Response rate and mean absolute relbias	-0.402(n = 57)	-0.552 (n = 43)
Response rate and mean absolute difference	-	-0.508 (n = 43)
Study-level correlations weighted by mean sample size		
Response rate and mean absolute relbias	-0.413 (n = 57)	-0.247 (n = 43)
Response rate and mean absolute difference	-	-0.208 (n = 43)
Estimate-level ICCs from multilevel model		
Absolute relbiases	0.164 (n = 955)	0.161 (n = 802)
Absolute differences	-	0.509 (n = 802)
Estimate-level R^2 from one-way ANOVA		
Absolute relbiases	0.221(n = 955)	0.211 (n = 802)
Absolute differences	-	0.395 (n = 802)

The results in Table 2.1 are important because responsive and adaptive designs work primarily at the study level. For example, case prioritization generally either increases the overall response propensities or reduces the variation in the propensities, and these are the two main study-level variables affecting the level on nonresponse bias. In addition, if a design succeeds in reducing the overall variation in the response propensities, this will tend to attenuate the correlations between the propensities and the survey variables across the board. At the extreme, if there is no variation in the response propensities, the correlation with all the survey variables will be zero and there won't be any nonresponse bias. The results in Table 2.1 seem to contradict the view that response rates don't matter. Nonresponse rates are clearly an imperfect proxy for nonresponse bias, but they are an important predictor of the average level of bias in the estimates from a survey.

2.3 Experimental evaluations of responsive and adaptive designs

How well do responsive and adaptive designs achieve their goals? At the outset, I should note that our expectations shouldn't be too high. As we noted in an earlier paper (Tourangeau et al., 2017, page 208), these designs "represent an attempt to do more with less or at least to do as much as possible with less" in an increasingly survey unfavorable environment. To date, studies have used four basic strategies to achieve one or more statistical goals – multi-phase designs (like the one described by Groves and Heeringa, 2006), other types of case prioritization (in which different cases are slated to receive different levels of effort), adaptive contact strategies (changing the timing of contact attempts based on propensity models to maximize the chances of making contact), and tailoring of the field work or mode of data collection based on what is known about the cases before they are fielded. I briefly review some of the major efforts to evaluate each of these approaches.

Multi-phase designs and case prioritization. Peytchev, Baxter and Carley-Baxter (2009) report another study that, like Cycle 6 of the NSFG, employed a multi-phase design. They conducted a telephone study with two phases. The second phase used a much shorter questionnaire and offered a larger incentive than the first. Cases received up to twenty calls during Phase 1, with some cases getting even more. Overall, this phase produced a response rate of 28.5 percent. In Phase 2, the researchers subsampled the remaining nonrespondents, shortened the questionnaire from 30 to 14 minutes, gave a prepaid incentive of $5, and offered a conditional incentive of $20. (Phase 1 had offered only conditional incentives.) Phase 2 produced a response rate of 9.8 percent (or 35.5 percent overall). The evaluation of the design was based two sets of comparisons: Peytchev and his colleagues compared early and late respondents from Phase 1 and they compared Phase 1 to Phase 2 respondents. They reasoned that the late respondents (interviewed after at least six call attempts) from Phase 1 were unlikely to differ on the key study variables – reported crime victimizations of various sorts – from the early respondents (interviewed in five or fewer attempts) because they were recruited via the same protocol. The results indicated that the addition of the late Phase 1 respondents did not significantly change the estimates. In contrast, the authors believed the Phase 2 respondents were likely to differ from the Phase 1 respondents, because the changes in protocol would attract different types of respondents. There was some support for this line of argument for males.

The Phase 1 male respondents were more likely to report victimizations than the Phase 2 male respondents, with significant differences on four of six victimization rates. However, there was less evidence that the change in protocol in Phase 2 affected the estimates for females. In addition, even within the Phase 1 sample, there were differences between male cases who never refused and those who were converted after refusing. Like the Phase 2 male respondents, the converted Phase 1 male refusals also showed significantly lower victimization rates on four of six key estimates. This suggests that the refusal conversion protocols changed the make-up of the Phase 1 sample and did not just bring in more of the same type of respondents.

Peytchev, Riley, Rosen, Murphy and Lindblad (2010) report a study that tailored the data collection protocol for different groups of cases from the outset. Their study involved a panel survey and the response propensities for each case was estimated using information from the prior round. Cases with low predicted response propensities were randomly assigned to an experimental or control treatment. For most of the data collection period, interviewers got a $10 bonus for each completed interview with one of the control cases, but $20 for each completed interview with one of the experimental cases. (During Phase 1, there was no bonus for control interviews and a $10 bonus for experimental interviews.) There was little difference in the final response rates for the two groups of cases (89.8 percent for the control cases versus 90.8 percent for the experimental cases) or in the average number of contact attempts per case (5.0 for the controls versus 4.9 for the experimental cases). Although the variance in the estimated response propensities was lower among the experimental cases, the estimated nonresponse biases (based on the correlations between the survey variables and the fitted response propensities) were higher.

Another set of experiments illustrates some of the practical difficulties with case prioritization. Wagner, West, Kirgis, Lepkowski, Axinn and Kruger Ndiaye (2012; see also Lepkowski, Mosher, Groves, West, Wagner and Gu (2013)) carried out 16 experiments over the course of Cycle 7 of the NSFG, which fielded 20 quarterly samples. The experiments examined the effectiveness of "assigning a random subset of active cases with specific characteristics to receive higher priority from the interviewers… The first objective of these experiments was to determine whether interviewers would respond to a request to prioritize particular cases" (Wagner et al., 2012, page 482). In only seven of the 16 experiments did the priority cases actually receive significantly more calls than the control cases, and only twice did this lead to a significant increase in response rates for the priority cases. Additional experiments attempted to shift the effort of NSFG interviewers from trying to complete main interviews to trying to complete screeners during one week of the field period. This intervention did lead to more screener calls than in prior or later weeks, but the impact on the number of *completed* screeners varied across quarters. In both cases, the efforts at case prioritization in Cycle 7 of the NSFG had some impact on what the interviewers did, but less impact on the intended survey outcomes, such as response rates.

Statistics Canada has also begun implementing responsive designs for its CATI surveys and carried out two experiments assessing these designs. Both experiments used three phases of data collection with case prioritization in one phase (Laflamme and Karaganis, 2010; Laflamme and St-Jean, 2011). In Phase 1, cases were categorized by response propensities; in Phase 2, cases were randomly assigned either to the

responsive collection condition (in which cases were assigned priorities and the high priority cases got more calls) or the control condition; and in Phase 3, all remaining cases got the same treatment. In Phase 2, the priority cases in the responsive collection group were apparently those with high predicted response propensities. The goal in Phase 3 was to equalize response propensities across key subgroups. Once again, the results indicated modest effects. The overall response rates were essentially unaffected by case prioritization. In one survey, the response rates were 74.0 percent for the control group versus 74.1 for the responsive collection group; in the other, the control group had a slightly higher response rate (73.0 versus 72.8 percent). This is a little surprising since the responsive collection targeted the easier cases in Phase 2. In addition, neither the new three-phase design nor the responsive collection protocol had a clear effect on the representativeness of the samples, but may have decreased the number of interviewer hours (see Table 2.2 in Laflamme and St-Jean (2011)). Still, reducing costs without reducing representativeness may represent a worthwhile, if modest, advance.

Adaptive contact strategies. Can survey managers improve the rate at which sample members are contacted by modelling the best time to contact them? Although many papers have explored optimal times for contacting sample members in surveys, few have examined whether these "optimal" call schedules produce gains empirically. Wagner (2013) is an exception. He reported five experiments that used models to predict whether a given sample household would be contacted on the next call attempt in each of four call "windows" (e.g., Tuesday through Thursday from 4 p.m. to 9 p.m.). Similar models were used in telephone (the Survey of Consumer Attitudes, or SCA) and face-to-face (Cycle 7 of the NSFG) surveys. The models were used to identify the best call window (the one with the highest probability of a contact) for each sample household. In the experimental groups, cases were moved to the top of the list for calling in that window (in the SCA) or field interviewers received that window as the recommended time to contact the household (in the NSFG).

Three experiments involved the SCA. In the first, the proportion of calls producing a contact was higher for the experimental cases than for the controls (12.0 percent versus 9.9 percent), but the strategy seemed to backfire for cases who had initially refused, with *lower* contact rates among the initial refusals in the experimental group. A second experiment varied the call window for experimental cases after an initial refusal but this strategy lowered the overall proportion of calls producing a contact. The final SCA experiment still found that the contact rate for refusal conversion calls was lower in the experimental group than in the control group. The results in the NSFG were also somewhat disappointing. The field interviewers apparently ignored the recommended call windows; only 23.6 percent of the experimental cases were contacted in the recommended window (versus 23.0 percent in the control group). We had a similar experience in our effort to get interviewers to follow an optimal route in their trips to the field (see Section 3.1 below).

Tailored field work. Luiten and Schouten (2013) report an experiment that tailored the data collection approach to different subgroups in the Dutch Survey of Consumer Sentiments (SCS). The goal was to equalize response propensities across the subgroups. The SCS consists of repeated cross-sectional surveys and, based on earlier rounds, Luiten and Schouten fit contact and cooperation propensity models based on

demographic characteristics of the sample members; these variables were available for the entire sample from the population registry. There were two phases of data collection. In the initial phase, cases with lowest estimated cooperation propensities were sent a mail questionnaire; those with the highest estimated propensities were invited to complete a web survey; and those in the middle were given a choice between mail and web. The second phase consisted of following up nonrespondents by telephone. Cases in different contact propensity quartiles were assigned to different call schedules. Those with the highest estimated contact propensities were fielded later in the field period and called during the day; those in the second highest quartile were called twice at night and then switched to a schedule alternating daytime and nighttime calls; and those in the lowest two contact propensity quartiles were called on every shift of every day. Finally, the best telephone interviewers were assigned to the cases with the lowest estimated cooperation propensities and the worst telephone interviewers were assigned to the cases with the highest estimated cooperation propensities. The control group for the experiment was the regular SCS, which is a CATI-only survey.

Although the adaptive field work group had only a slightly higher response rate than the regular SCS (63.8 percent versus 62.8 percent, a non-significant difference), the representativeness of the experimental sample, as measured by the R-indicator, was significantly higher than that of the control sample. (The R-indicator, introduced by Schouten, Cobben and Bethlehem (2009), is based on the variation in the estimated response propensities. A higher number indicates less variation and therefore a more representative sample.) Table 2.2 below shows that the adaptive field work did lower the variation in both contact and cooperation rates. Across contact propensity quartiles, the contact rates ranged from 84.2 percent to 96.9 percent in the regular SCS; in the experimental sample, the range was from 87.1 to 95.3. The adaptive design also lowered variation in the cooperation rates. Still, the costs for the adaptive design were marginally higher than those of the SCS and the overall cooperation rate was significantly *lower* in the experimental sample. Unfortunately, as this study illustrates, reducing the variability in the response propensities often means not trying as hard to get the easiest cases and this may lower the overall response rate.

Table 2.2
Contact and cooperation rates, by propensity quartile groups

	Contact rates	
Contact propensity quartile	**Experimental**	**Control**
Lowest Contact Propensity	87.1	84.2
Second Lowest Contact Propensity	96.6	94.5
Second Highest Contact Propensity	93.7	95.7
Highest Contact Propensity	95.3	96.9
	Cooperation rates	
Cooperation propensity quartile	**Experimental**	**Control**
Lowest Cooperation Propensity	65.1	62.7
Second Lowest Cooperation Propensity	71.4	68.4
Second Highest Cooperation Propensity	72.8	75.3
Highest Cooperation Propensity	74.7	79.2

Source: Tourangeau et al. (2017); data from Luiten and Schouten (2013).

2.4 Simulation studies

Besides the experiments discussed in the previous section, three additional studies have used simulations to explore the properties of responsive and adaptive designs.

Stopping rules. Lundquist and Särndal (2013) used data from the 2009 Swedish Living Conditions Survey (LCS) to explore the impact of various "stopping rules", rules for ending data collection. The LCS follows a two-phase data collection strategy, with up to 20 telephone contact attempts in the first phase of data collection followed by ten more in the second phase. They noted that continuing to follow the same data collection protocol "will produce very little change in the estimates beyond a certain 'stability point' reached quite early in the data collection" (page 561). This is quite similar to Groves and Heeringa's (2006) notion of "phase capacity", or the point at which a given data collection protocol begins to achieve diminishing (or vanishing) returns. Sturgis, Williams, Brunton-Smith and Moore (2017) present results suggesting that this stability point may be reached quite early during the field period. They examined estimates derived from 541 questions from six face-to-face surveys in the U.K. They found that the expected proportions were, on average, only 1.6 percent from the final estimate after a single contact attempt and were off by only 0.4 percent after five attempts. These results suggest that, from the vantage point of reducing bias, a lot of field effort is wasted.

Lundquist and Särndal show that the estimated nonresponse bias (based on three variables available for both respondents and nonrespondents from the Swedish population register) in the LCS was lowest after five to ten call attempts and actually got progressively worse thereafter. The second phase of data collection, which increased the response rate from 60.4 percent to 67.4 percent, made the nonresponse biases worse for two of the three register variables. They examined three alternatives to continuing the same protocol up to 30 attempts. They divided the sample into eight subgroups based on education, property ownership, and national origin. Under the first alternative response rates for each of eight the subgroups would be checked at call 12 of the initial phase of data collection and again at call 2 of the second phase; data collection would end for subgroups with response rates of 65 percent or better at these points. This strategy would have yielded a lower response rate (63.9 percent) than the actual protocol but a sample that was more closely aligned with the population on eight demographic characteristics. The second alternative they examined would have ended data collection for a subgroup as soon as its response rate reached 60 percent and the third alternative, as soon as the subgroup response rate reached 50 percent. The 50 percent strategy would have produced the most balanced sample of all and would have reduced the total number of call attempts by more than a third. In part, this strategy worked so well because it would have lowered the response rates in the high propensity subgroups so they were closer to those in the low propensity subgroups. As in the study by Peytchev and colleagues (Peytchev et al., 2009), continuing with the same data collection protocol seemed to do little improve the representativeness of the sample, and may in fact have reduced it.

In a related effort, in 2017, the Medical Expenditure Panel Survey (MEPS) used a stopping rule based on a propensity model. MEPS is a rotating panel study. Each year a new panel of about 10,000 addresses

is selected from a sample of households that completed National Health Interview Survey the previous year. Sample households were asked to complete two MEPS interviews in their first year, two in the second, and a fifth in the third year. The survey is continuous, with interviews conducted throughout the year. The stopping rules were applied in two stages in the first half of 2017: first to cases in their third round (a relatively soft start, since most Round 3 interviews were scheduled by telephone and most respondents were cooperative, having already participated twice), and then to Round 1 cases. Interviewers are often reluctant to comply with directions to stop contacting a case after a specific number of attempts. The MEPS approach was to remove low propensity cases with too many attempts – generally six – from the interviewer assignment and have a supervisor review them. Supervisor could move a case back into the interviewer's assignment if there was some reason to believe the case might be completed, but most of the time these cases were closed out (Hubbard, 2018). Overall, implementing the stopping rule reduced the number of in-person attempts by 8,500, producing a large saving in field costs.

Different case prioritization strategies. In a later paper, Särndal and Lundquist (2014b) simulated the effects of two methods for equalizing response propensities across cases, using data from the Living Conditions Survey and the Party Preference Survey. Under the first method (the *threshold* method), no further follow-up attempts are made to cases whose response propensities have reached some threshold (lower than the overall target response rate). This is similar to the strategies examined in their earlier paper (Lundquist and Särndal, 2013). Under the other method (the *equal proportions* method), at various points during the field period (e.g., after three, six, or nine call attempts), the portion of the sample with the highest response propensities is set aside and field work continues only for the remaining cases. In both surveys, both methods for equalizing the response propensities reduced the distance between the respondents and the full sample on a set of auxiliary variables, as compared to continuing to field all remaining nonrespondents, as was done in the actual surveys. Another conclusion from this study is that calibrating the sample using the auxiliary variables removed some of the nonresponse bias, but that bias was reduced even further when the set of respondents was more closely aligned with the population in the first place. This is an important finding, since the same variables available for fitting propensity models are also available for post-survey adjustments, and it is not clear whether equalizing response rates (or response propensities) during data collection is more effective than simply adjusting the case weights afterwards. Särndal and Lundquist (2014b) find gains for both.

Beaumont, Bocci and Haziza (2014) report another simulation study that examines the impact of case prioritization. They contrasted four strategies: 1) *constant effort* (no case prioritization); 2) *optimal effort* (by reducing calls to members of groups approaching their target response rate); 3) *equalizing response rates* across groups (by concentrating calls on low response propensity groups); and *maximizing the overall response rate* (by concentrating calls on high propensity groups). The simulations by Beaumont and his colleagues assumed three different scenarios – uniform response propensities, uniform response propensities within groups, and response propensities that are highly $(r = 0.67)$ correlated with the survey variable of interest. (In addition, the simulation assumed that the sample consisted of three

subgroups, that calls yielding an interview were 25 times more expensive than ones that didn't, that calls to a case were capped at 25, and the survey had a fixed data collection budget.)

The simulations supported three major conclusions. First, when response propensities are constant overall or constant within each group, all the effort strategies produce unbiased estimates, but when the propensities were strongly related to the survey variable, all of them produced bias. Second, neither the R-indicator nor the nonresponse rate was a good indicator of nonresponse bias or nonresponse variance. Finally, when response propensities were known, the optimal effort strategy produced somewhat lower root mean square error than the other strategies (see Table 2.2 in Beaumont et al. (2014)) and the strategy that attempted to maximize response rates produced the worst. The optimal effort strategy resembles the approaches explored by Lundquist and Särndal (2013). Of course, a practical difficulty is that response propensities are not known with real surveys, and they may not be accurately estimated from the available auxiliary variables.

2.5 Summary

Table 2.3 summarizes the results from the experimental and simulation studies. In general, they show how hard it is to raise response rates in the current environment. For example, only two of the 16 experiments described by Wagner and his colleagues significantly raised response rates in the NSFG (Wagner et al., 2012). Some studies (e.g., Luiten and Schouten, 2013) demonstrate reductions in variation in response rates across subgroups of the sample, although in one study (Peytchev et al., 2010) this apparent reduction in the variation in estimated response propensities appeared to *increase* nonresponse bias rather than reduce it. Laflamme and St-Jean (2011) reported that responsive design reduced costs relative to the standard protocol, but Luiten and Schouten (2013) reported that an adaptive design increased the costs per case. Across all the studies (including Cycle 6 of the NSFG), then, responsive and adaptive designs appeared to produce some gains in sample representativeness, but had little effect on overall response rates or overall costs.

Several non-experimental studies come to similar conclusions. These studies compare the final survey estimates with those that would have been obtained without the final phase of data collection, when a major change in the data collection protocol was introduced. For example, Groves and his colleagues (Groves et al., 2005) showed that the final phase of data collection in Cycle 6 of the NSFG, which boosted the overall response rate from 64 to 80 percent, also decreased variation in the response rates across subgroups (see also Axinn, Link and Groves, 2011). This is similar to the experimental results reported by Peytchev, Baxter and Carley-Baxter (2009) who found that major changes in protocol (larger incentives and a shorter questionnaire) produced changes in the study estimates, at least for males. However, the changes were generally small – less than two percentage points.

Table 2.3
Selected study characteristics and outcomes, by study

Experimental Study	Statistical Goal	Intervention	Results
Peytchev et al. (2010)	Equalize response propensities	Bonus for interviewers for completing high priority cases	• Variance in response propensities lower in experimental group • Response rate 1.5% higher in experimental group • Estimated bias *higher* in experimental group
Wagner et al. (2012)	Increase response rates, improve representativeness	Case prioritization Screener week	• Significantly increased number of calls to priority cases in seven of 16 experiments • Significantly increased response rate in two experiments • Increased number of screening calls
Laflamme and St-Jean (2011)	Increase response rates (Phase 2), equalize response propensities (Phase 3)	Categorization and prioritization of cases	• Less variance in response propensities in experimental group • Response rate 1.5% higher in experimental group
Wagner (2013)	Increase contact rate per call	Models used to assign cases to optimal call window	SCA • Contact rate improved (12.0 vs. 9.9 percent) • No change in response rate NSFG • Interviewers did not follow recommended call window
Luiten and Schouten (2013)	Equalize response propensities	Initial mode (mail versus Web) varied by propensity quartile Hard cases assigned to best telephone interviewers, easiest to worst telephone interviewers	• Lower cooperation rate in adaptive group • R-indicator significantly improved in adaptive group • Reduced variation in contact and cooperation rates in adaptive group • No significant difference in costs or response rates
Simulation Study	**Statistical Goal**	**Intervention**	**Results**
Lundquist and Särndal (2013)	Increase sample balance, reduce nonresponse bias	Stopping data collection for a subgroup once a target rate achieved for that subgroup	• Lowest response rate threshold produced the highest balance • Lowest threshold also achieved lowest nonresponse bias (on three registry variables)
Särndal and Lundquist (2014a, b)	Increase sample balance, reduce nonresponse bias	12 stopping rules	• Lowest response rate threshold again produced the highest balance • Lowest threshold also achieved lowest nonresponse bias on three registry variables • Both balance in data collection and calibration reduce nonresponse bias
Beaumont, Bocci and Haziza (2014)	Optimal effort, equalize response rates, maximize overall response rate	Four case prioritization strategies (constant effort, reduce effort for groups approaching target response rate, prioritize low propensity cases, prioritize high propensity cases)	• With uniform response propensities, all four strategies yield unbiased estimates • When response propensities strongly related to survey variables, all strategies produce biased estimates • With known propensities, optimal strategy yields best root mean square error (RMSE); maximizing the response rate, the worst

3. Optimal routing

Case prioritization, like adaptive design more generally, is an example of an intervention in the data collection process intended to reduce error, costs, or both. In the next two sections, we examine two other interventions designed to improve data collection outcomes – optimal routing and rapid feedback to

interviewers. The first uses a variation on case prioritization; the second focuses on reduction of measurement error.

Prioritizing high-value cases. One problem with the existing studies on case prioritization is that they have all used estimated response propensities as the basis for prioritization. The response propensity may be a useful summary of the variables used to model the propensity but may not fully reflect the researchers' priorities. In an earlier paper (Tourangeau et al., 2017), we proposed a different basis for case prioritization. Under our scheme, the cases receiving the highest priority should be the ones with the highest ratio of anticipated value to anticipated cost:

$$B_i = \frac{\hat{\rho}_i W_i V_i}{C_i} \tag{3.1}$$

where B_i represents the benefit-to-cost ratio for case i; the numerator is the product of the case's estimated response propensity $(\hat{\rho}_i)$, its weight (W_i), and some measure of its value for the research (V_i); and the denominator represents the likely cost of completing the case (C_i). For example, the value assigned to a member of a rare subgroup may be higher than that assigned to a member of a larger group. Or the value of a case may be an estimate of its impact on reducing the distance between the current sample from a vector of population benchmarks. Because it includes the estimated propensity and the weight in the numerator, the scheme in (3.1) may result in giving priority to "easy" cases or give lower priority to cases from oversampled subgroups, which would have lower weights. Thus, a lot hinges on how the value of a case is assessed. We attempted to apply a version of (3.1) in conducting a pilot test of a strategy that we call *optimal routing*.

The pilot test. Our pilot test was done as part of the Population Assessment of Tobacco and Health (PATH) Reliability and Validity Study (the PATH-RV Study; Tourangeau, Yan, Sun, Hyland and Stanton, 2019), a study designed to assess the reliability and validity of answers to the Wave 4 PATH Study questionnaires. (The PATH Study is a major longitudinal study of tobacco use, and the study sponsors wanted to be sure the questions yielded reliable responses.) In the PATH-RV study, a sample of 524 respondents completed the PATH questions twice, roughly two weeks apart. There were two questionnaires, one for adults (18 years old and older) and one for youths (12 to 17 years old). Given the aims of the reliability study, we deemed youth cases to be twice as valuable as adult cases (because youths were rarer and harder to interview than adults) and reinterviews 1.5 times more valuable than initial interviews. Thus, an initial interview with an adult was worth a value of 1; an adult reinterview, 1.5; an initial youth interview, 2; and a youth reinterview, 3. We used these values in the place of V_i in (3.1). Because the sample for the PATH-RV study was nearly equal probability, we ignored the weights. However, we did incorporate an estimate of the likelihood that the case would cooperate on the next contact attempt. We also developed a program that calculated an optimal route for contacting a set of cases on a given day, partly in an effort to minimize travel and interviewer time – that is, to minimize C_i in (3.1).

The system we developed had two components. The first one estimated the likelihood that each remaining case would cooperate on the next contact attempt. The models for this first component used sociodemographic information from the Census Planning Database for block groups and the history of previous contact attempts whenever at least one contact attempt was available. For cases with no prior contact attempts, we used a logistic regression model to estimate the cooperation propensity; for cases with prior contact attempts, we used a proportional-hazards Cox regression model. The second component was a routing system that reviewed respondent-level information and produced an interviewer's schedule for a given day. The set of cases in the day's assignment reflected the anticipated value of the cases. All the sample cases were geolocated, allowing us to estimate the travel time between each pair of cases for a given hour of the day. The routing system took as input the feasible tasks for a given case (e.g., it did not schedule a reinterview until the initial interview had been completed), along with the case's geographical location, estimated duration of the task, case value, and response propensity. It then computed the shortest driving route with the highest possible expected value and selected a set of tasks that could fit in a working day for a given interviewer. The route delivered to interviewers included the sequence of cases and tasks that we expected interviewers to attempt. It took appointments into account, and the route was constructed to ensure that interviewer could arrive at their appointments on time.

The experimental design. We conducted an experiment that compared interviewer performance on "treatment" days when we gave them the list of cases to try to get along with a suggested route to follow in pursuing those cases with "control" days when we gave them no special instructions about which cases to work or how to work them. The data collection for the experiment took place between October and December, 2017.

Before the start of data collection, interviewers selected at least six days during which they would work only on the PATH-RV study. We then randomly allocated three of those days to the control arm of the experiment and three to the treatment arm. On control days, we sent interviewers an email in which we asked them to "use their best judgement on how to contact" their caseload. On treatment days, we sent them an email that included a list of cases that we wanted them to work and the route they were to follow. Interviewers were told to follow our recommendations "if at all possible". During the training sessions and in the email accompanying the selected route, we discouraged deviations from the instructions, but allowed them if the interviewers judged them necessary to account for unforeseeable events, such as traffic accidents.

Fifty-three interviewers participated in the experiment. Changes to the days the interviewers worked on our study, together with the depletion of the pool of open cases in the final days of the study, produced a reduction in the number of treatment and control days actually available for the interviewers. Ultimately, we had a total of 220 observations.

Interviewer compliance and interviewer efficiency. Did the interviewers follow the instructions we sent them in the treatment email? Well, they did some of the time. There was an average overlap of 62 percent between the cases we recommended for a given treatment day and the cases the interviewers

actually worked that day. What is particularly striking is that there was, on average, a 52 percent overlap on the cases selected by our model and the cases selected by the interviewers on the control days, when we didn't give them any instructions. This small difference between the treatment and control days partly reflects the limited number of cases that could be worked on any given day. As a result, the decisions that interviewers would have made on their own were often close to what we thought would have been optimal, putting a low ceiling on the possible impact of the treatment.

Still, there was only moderate compliance with instructions by the interviewers. A Census Bureau test had similar results. The test was done in eight areas in Philadelphia, Pennsylvania (Walejko and Miller, 2015). In some areas, interviewers were assigned seven high priority cases each day; these high priority cases were those deemed most likely be interviewed on the next contact attempt, according to response propensity models. As with our experiment, interviewer compliance was an issue. As Walejko and Miller (2015) put it: "The ability of response propensity models to identify promising cases for daily contact, however, remains unclear after this pilot test because interviewers did not dutifully work priority cases."

Was there any sign in our study that the optimal routing treatment improved interviewer efficiency? We examined five outcomes of interest:

1) The number of miles interviewers traveled;
2) The hours they spent;
3) The number of contacts per completed interview;
4) The number of completed cases; and
5) The average value of the cases completed.

The first two variables reflect the impact of the treatment on the costs of collection. We also wanted to assess whether our routing system reduced the number of contact attempts needed to complete a case – that is, whether it made the interviewers more productive. Similarly, we examined whether the treatment increased the number of completes and whether the completed cases had higher values on average on the treatment days than on the control days. Our analyses of the effects of the treatment are shown in Table 3.1. The models include random effects for each interviewer and pool the effect of the treatment across interviewers. The top two panels show the estimates for the intercept and treatment effects under an intent-to-treat model (ignoring whether the interviewers actually followed our instructions), and the second panel incorporates a measure of the interviewers' compliance with the instructions. None of the outcome measures shows a significant treatment effect, although there were significant compliance main effect for miles and contacts – interviewers traveled fewer miles and made fewer contacts when they did what we suggested (whether we conveyed those suggestions to them or not). Although there was a treatment by compliance interaction effect on contacts, the net effect of the treatments seems to have been negative.

Table 3.1
Estimated intercepts and effects (and standard errors), by outcome and model

	Miles	Hours	Contacts	Completes	Value
Intent-to-treat					
Intercept	76.8 (7.0)	5.26 (0.37)	5.40 (0.53)	0.81 (0.17)	1.73 (0.26)
Treatment	8.06 (5.6)	0.04 (0.27)	-0.28 (0.48)	-0.15 (0.16)	-0.24 (0.30)
Incorporating compliance					
Intercept	89.2 (8.99)	5.69 (0.46)	7.66 (0.62)	0.98 (0.21)	1.84 (0.40)
Treatment	0.30 (9.87)	-0.30 (0.48)	-1.35 (0.79)	0.15 (0.28)	0.25 (0.53)
Compliance	-28.9 (13.4)	-1.01 (0.65)	-5.32 (1.06)	-0.03 (0.37)	-0.28 (0.70)
Treatment x compliance	20.5 (17.4)	0.86 (0.85)	3.07 (1.38)	-0.55 (0.48)	-0.87 (0.92)

Note: Results based on 53 interviewers and 220 total observations.

Interviewer reactions. Debriefings with the interviewers revealed some of the reasons for their relatively low levels of compliance with our recommendations. Although the interviewers were generally positive about the routing system, they had several reservations about it. The behavior of interviewers reflects the goal of getting completed interviews, but their implicit assumption is that all completes are equally valuable. However, our routing system reflected a specific definition of the expected value of a case and also an estimate of its cost. As a result, it sometimes omitted cases that were close to the households on the recommended route. Interviewers indicated that a priority list or a scoring of the cases by their value would have made the decisions of the automatic system more comprehensible and also would have allowed them to incorporate those values into their own workday planning. In addition, interviewers sometimes disagreed with the suggested routes because of circumstances that could not be observed by our routing system. With any adaptive design strategy (or, more generally, with any planning system), there is the risk of missing some useful information and this may undercut compliance.

The debriefing also called attention to some of the assumptions embedded in the model. For instance, we established a single time window for all interviewers as the most likely time they would be working. This allowed us to account for daily traffic patterns in our recommendations. But a different route might have been better than the one we recommended for a different time of day when traffic was lighter or heavier. All the interviewers who took part in the experiment were experienced field interviewers, and some reported they felt that detailed routing instructions were tantamount to discounting their abilities and experience. In their opinion, the system might be a good tool for novice interviewers, but, for them, it signaled a lack of confidence on the part of the survey managers. Finally, they all reported that one reason they worked as field interviewers was being able to plan their own workday. Many of these same factors doubtless played a role in the limited success of the attempts by Wagner and his colleagues (see Section 2.3) and the Census test to change interviewer behavior.

Despite these obstacles to compliance, research has shown that interviewers are sensitive to incentives. Tourangeau, Kreuter and Eckman (2012) demonstrated that interviewers in a telephone study completed more screeners when they were given a bonus for each screener they completed and they completed more main interviews when they were given a bonus for each completed main interview. Perhaps similar incentives could be used to encourage interviewers to complete high priority cases or to minimize travel

time. For example, interviewers could receive a small bonus for every high priority case they contact. Clearly, we need to figure out how to get interviews to follow instructions if our interventions are going to have any impact.

Other studies of interviewer travel. More recently, Wagner and Olson (2018) carried out an extensive analysis of interviewer travel in two face-to-face surveys, the National Survey of Family Growth (NSFG) and the Health and Retirement Survey (HRS). Both surveys feature national area probability samples and the Survey Research Center at the University of Michigan carries out the field work for both. The surveys have different target populations – people from 15-44 years old in the NSFG and from 51-56 years old in the HRS. The authors examined how far interviewers travelled and how many sample areas they visited on each day they worked. In both studies, interviewers visited about two areas, on average, on each day they worked but they travelled about 30 miles more in average in the NSFG than in the HRS (85.4 versus 53.4). Wagner and Olson found that travelling to more areas was associated with more contact attempts, but with fewer contacts made and fewer interviews completed (see their Table 4.1). Although theirs is an observational study and not an experiment, it is consistent with the results of our pilot study; more travel seems to reduce the number of contacts made and interviews completed. However, the causal direction of this finding is quite ambiguous. It could be that travel time reduces the time interviewers have left to contact and interview sample cases, but it also could be that interviewers keep going when their contact attempts don't yield a positive outcome, moving on to different sample areas.

4. Rapid CARI feedback

Interviewers can contribute in several ways to the total error of a survey estimate, affecting coverage, nonresponse, and measurement errors (Schaeffer, Dykema and Maynard, 2010; West and Blom, 2017). There can be complex interactions among these different interviewer-related error sources. For example, there may be a tradeoff between coverage and nonresponse errors (Tourangeau et al., 2012); in our study, the interviewers with the highest response rates also found the fewest eligible households. In a series of papers, West and his colleagues (West and Olson, 2010; West, Kreuter and Jaenichen, 2013; West, Conrad, Kreuter and Mittereder, 2018) have shown that different interviewers may elicit different answers because of differences in the respondents they recruit (e.g., some interviewers may be better than others at recruiting older respondents) but also because of differences in their levels of measurement error. As anyone who has ever listened to CARI recordings can testify, interviewers do not always stick to the script and their improvisations can sometimes elicit poor quality responses.

Pilot study. Having listened to recordings of field interviewers as part of the field test for a major national study, we designed an experiment to test the hypothesis that providing timely feedback to interviewers about their reading of the questions would improve the quality of the answers they elicited. (At the client's request, we do not divulge the name of the study.) This particular survey was a good test bed for assessing the effects of rapid feedback because the interviewers administered a short screening questionnaire to a household informant and then similar questions were administered to each sample

member via audio computer-assisted self-interviewing (ACASI). As a result, we could compare the screening data collected by each interviewer with a "gold standard" for several of the key items in the survey. Of course, the ACASI data are not error-free, but we regarded them as less error-prone than the screener data for two reasons: Each person reported for himself or herself whereas the screener was administered to a single household informant; and the questions were self-administered rather than administered by an interviewer and self-administration was likely to reduce any social desirability bias in the responses.

The experiment included 291 interviewers. Half were assigned to receive rapid feedback and half were assigned to the control group. Every fifth screener done by interviewers in the rapid feedback group was CARI-coded to identify departures from standardized interviewing. Figures 4.1 displays the questions coders answered for each screening interview. After a screener was coded, interviewers (and their supervisors) were sent a report with their performance and a link to the question recordings. For their first coded screener, interviewers were instructed to schedule a feedback session with a central office "mentor", who reviewed the results and provided guidance for improvement. For their second coded screener, interviewers were sent only the report and a link to the recordings. For subsequent screeners, interviewers were only instructed to schedule a feedback session with their mentor if the coding identified problems; otherwise, they were only sent the written report.

Figure 4.1 Coding questions for rapid feedback pilot study. The questions were repeated for each member of the household.

Q1. How clearly can you hear the interviewer on this recording? [HEARINT]
- Very clearly (4)
- Somewhat clearly (3)
- Not very clearly (2)
- Cannot hear the interviewer (1)

Q2. How clearly can you hear the respondent on this recording? [HEARRESP]
- Very clearly (4)
- Somewhat clearly (3)
- Not very clearly (2)
- Cannot hear the interviewer (1)

Q3. Did the interviewer read the question exactly as worded? [EXWORD]
- Yes (1)
- No (2)

Q4. [IF NO TO Q3] How did the interviewer change the wording of the question? Pick all that apply
- Did not read lead-in or introductory text before the question [NOINTRO]
- Did not read "Please look at this picture" [NOPIC]
- Did not read "Please look at this list" [NOLIST]
- Did not read all brand names or product examples [NONAMES]
- Did not read response options correctly [NORESP]
- Did not read "choose all that apply" [NOCHOOSE]
- Omitted, added, or changed other words within the question [NOREADOTH]

Q5. Did the interviewer correctly enter the respondent's answer? [ENTERANS]
- Yes (1)
- No (2)

The experiment was conducted from May to August, 2014, with 1,729 respondents interviewed by the feedback group and 1,717 interviewed by the control group.

To evaluate the effects of rapid feedback, we compared three variables derived from the screening items to the corresponding variables from the ACASI interviews. In principle, the two should match. Table 4.1 shows the proportion of respondents in the treatment and control groups who were classified the same way in the screener and the ACASI data. For all three, the match rate was significantly higher for respondents who were interviewed by interviewers getting feedback. (We used a Rao-Scott F test that took into account the clustering of the sample by areas. All three F -values were significant at $p < 0.01$.) Kappa values measuring the chance corrected agreement between screener and ACASI responses are substantially higher for interviewers in the rapid feedback group as well.

Table 4.1
Agreement between screener and ACASI responses, by condition and variable

	Rapid feedback	Control	Rao-Scott F value (1 and 230 df)
Composite			
% Agree	93%	88%	15.5***
Kappa	0.85	0.76	
Variable 1			
% Agree	95%	92%	8.8**
Kappa	0.89	0.83	
Variable 2			
% Agree	89%	85%	7.7**
Kappa	0.76	0.69	

** $p < 0.01$;

*** $p < 0.001$.

Note: The composite was a summary variable derived from variables 1 and 2.

MEPS study. Based on the success of this initial study, Edwards, Sun and Hubbard (2019) undertook a replication. In 2018, the Medical Expenditure Panel survey had implemented a major upgrade of the CAPI system and had simplified some sections of the questionnaire. Two question series were of particular interest because they were asked in all interviews, always recorded in CARI (almost all respondents gave consent to record), and were critical for producing data on the use and cost of health care services, key MEPS statistics. These were the questions on the use of calendars or other records of medical care during the interview and "provider probes", filter questions that prompt the respondent to recall services from various types of medical providers. The calendar series asked whether various records were available during the interview (e.g., a calendar with entries for medical visits, insurance statements, etc.), and who in the household was associated with each record type. The CAPI entry area for these items was a grid, with each household member listed on a row and each record type a column header. Interviewers could enter answers in any order, by person or by record type. The objective was to encourage respondents to bring records for all family members into the interview and to structure the questioning so that the records could be incorporated into the interview in any order. The provider probes

consisted of 15 questions about various kinds of health care providers. They were re-ordered in the technical upgrade to begin with three that accounted for the highest expenditures.

Audio-recordings of the calendar series and the provider probes series were reviewed by two behavior coders. The coding system allowed coders to call up specific interviewers or questions. Coders evaluated the overall quality of the interview and of each instance of asking the calendar series and the provider probes. The inter-coder agreement rate was 0.82. Verbal and written feedback was provided to the interviewer quickly (ideally within 72 hours of the interview). The next interview conducted by the interviewer was also coded, so that each interviewer had a pair of interviews in the data set, one just before and one just after feedback. Because the process was implemented in late fall only a subset (122) of the MEPS interviewers were available to participate in the study, resulting in 244 interviews in the data set. Data about the feedback interaction was also captured (such as whether the interviewer agreed with the feedback or asked for clarification). Again, we expected that interviewer behavior more consistent with the study protocol would be observed after feedback, both for overall interview quality and for each question series.

Table 4.2 shows the rapid feedback results for each question series. Interviewers maintained the meaning of the questions but did not follow the protocol exactly in the majority of instances $(n = 5{,}259)$, both before and after feedback. Still, question-asking behavior that followed the protocol exactly increased from 33.4 percent before feedback to 43.4 percent after feedback; failing to maintain the question meaning decreased from 9.8 percent before feedback to 3.7 percent after feedback. An F-test that took into account the clustering of the observations by interviewer found a significant overall difference between interviewer behavior before and after feedback, both overall $(F(2{,}118) = 3.86,\ p < 0.05)$ and for the provider probes $(F(2{,}118) = 5.71,\ p < 0.01)$. The differences for the calendar series was in the same direction but not statistically significant. These results, like those of the pilot study, indicate that rapid feedback to the interviewers can lead to marked improvements in how they administer the questions.

Table 4.2
Interviewer behaviors, before and after feedback, by question series

Interviewer Behavior	Calendar series		Provider probes		Both series	
	Before	After	Before	After	Before	After
Followed protocol exactly	18.6%	27.9%	43.3%	51.7%	33.4%	43.4%
Maintained meaning but did not follow protocol exactly	68.9%	65.1%	48.7%	46.4%	56.8%	52.9%
Did not maintain meaning	12.5%	7.0%	8.0%	2.0%	9.8%	3.7%
Total	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
n	1,240	759	1,832	1,428	3,072	2,187

5. Conclusions

The three main methods reviewed here have a mixed record of success. What lessons can we draw from these efforts to substitute data for intuition in the management of surveys?

The literature on responsive and adaptive design leads to several conclusions. First, it is important to clarify the statistical goals for the design at the outset of the survey and to monitor measures of quality related to these goals. Different strategies serve different goals. For example, equalizing response propensities may reduce nonresponse bias at the expense of a smaller sample size and increased sampling variance. It is essential to acknowledge such tradeoffs. Second, both the overall response rate and the variation in response propensities contribute to the average nonresponse bias. As a result, no single indicator gives a complete picture of the risk of error in a survey and survey managers should monitor multiple indicators, including changes in a set of key survey estimates. Advances in "dashboard" design (Mohadjer and Edwards, 2018) make it easier for central office staff and field supervisors to monitor a large number of indicators of how the field work is going. Third, simply continuing a given data collection protocol may not change the estimates much (Sturgis et al., 2017) and, in some cases, may decrease the representativeness of the sample (Lundquist and Särndal, 2013; Särndal and Lundquist, 2014). Under a given data collection protocol, the respondents recruited late in the field period are not likely to differ much from the ones recruited earlier. The sample will continue to overrepresent the cases with higher propensities under that protocol. To change the mix of respondents – and to improve the overall representativeness of the sample – may require major changes in the data collection protocol, such as much larger incentives, a switch to a different mode of data collection, or a much shorter questionnaire. These strategies all have their drawbacks, leading to the conclusion that sometimes the best strategy is just to cease further efforts by imposing stopping rules. Continuing to pursue cases with very low response propensities to respond is a formula for driving up costs without really improving the statistical properties of the final estimates.

Both the literature on responsive and adaptive designs and the study on case prioritization and optimal routing discussed in Section 3 above indicate that one factor limiting the effectiveness of central office interventions on field work is resistance by the interviewers. We need more research on how to improve interviewer compliance and on the impact of closer monitoring (or larger incentives) to ensure interviewers implement the desired changes in protocol. The studies on rapid feedback to the interviewers are encouraging in this regard. Both studies I reviewed in Section 4 indicate that when interviewers are given timely feedback on their administration of the questions they do a better job, and this reduces the level of measurement error in the answers they elicit.

One thing is certain. In an increasingly difficult climate for surveys, efforts to improve the management of surveys and to apply as much as science as possible in that endeavor will surely continue.

Acknowledgements

This paper was written in response to my receiving the Waksberg Award. I am grateful to Brad Edwards, Gonzalo Rivero and Tammy Cook for their helpful suggestions on this paper; to Aaron Maitland and Gonzalo Rivero for their help in designing and carrying out some of the studies described here; and to Statistics Canada for giving me the award.

References

Atrostic, B.K., Bates, N., Burt, G. and Silberstein, A. (2001). Nonresponse in U.S. government household surveys: Consistent measures, recent trends, and new insights. *Journal of Official Statistics*, 17, 209-226.

Axinn, W.G., Link, C.E. and Groves, R.M. (2011). Responsive survey design, demographic data collection, and models of demographic behavior. *Demography*, 48, 1127-1149.

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

Bethlehem, J.G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, 4, 251-260.

Brick, J.M., and Tourangeau, R. (2017). Responsive survey designs for reducing nonresponse bias. *Journal of Official Statistics*, 33, 735-752.

Brick, J.M., and Williams, D. (2013). Explaining rising nonresponse rates in cross-sectional surveys. *The ANNALS of the American Academy of Political and Social Science*, 645, 36-59.

Calinescu, M., Bhulai, S. and Schouten, B. (2013). Optimal resource allocation in survey designs. *European Journal of Operational Research*, 226, 115-121.

Chun, A.Y., Heeringa, S.G. and Schouten, B. (2018). Responsive and adaptive design for survey optimization. *Journal of Official Statistics*, 34, 581-597.

de Leeuw, E.D., and de Heer W. (2002). Trends in household survey nonresponse: A longitudinal and international comparison. In *Survey Nonresponse*, (Eds., R.M. Groves, D. Dillman, J.L. Eltinge and R.J.A. Little), New York: John Wiley & Sons, Inc, 41-54.

Edwards, B., Sun, H. and Hubbard, R. (2019). Behavior change techniques for reducing interviewer contributions to total survey error. Unpublished manuscript.

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 (Statistics in Society)*, 169, 439-457.

Groves, R.M., and Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias. *Public Opinion Quarterly*, 72, 167-189.

Groves, R.M., Singer, E. and Corning, A. (2000). Leverage-saliency theory of survey participation: Description and an illustration. *Public Opinion Quarterly*, 64, 299-308.

Groves, R.M., Benson, G., Mosher, W.D., Rosenbaum, J., Granda, P., Axinn, W., Lepkowski, J. and Chandra, A. (2005). *Plan and Operation of Cycle 6 of the National Survey of Family Growth*. Hyattsville: National Center for Health Statistics.

Hansen, M.H., and Hurwitz, W.N. (1946). The problem of nonresponse in sample surveys. *Journal of the American Statistical Association*, 41, 517-529.

Hicks, W.D., Edwards, B., Tourangeau, K., McBride, B., Harris-Kojetin, L.D. and Moss, A.J. (2010). Using CARI tools to understand measurement error. *Public Opinion Quarterly*, 74, 985-1003.

Hubbard, R. (2018). That wasn't part of the plan! Reducing effort through stopping rules to place CAPI cases on hold and work plans to set them free. Paper presented at the Annual Conference of the American Association for Public Opinion Research, Denver, Colorado, May 17, 2018.

Laflamme, F., and Karaganis, M. (2010). Development and implementation of responsive design for CATI surveys at Statistics Canada. Paper presented at the European Quality Conference, Helsinki, Finland.

Laflamme, F., and St-Jean, H. (2011). Highlights and lessons from the first two pilots of responsive collection design for CATI surveys. In *Proceedings of the Section on Survey Research Methods*, Alexandria, VA: American Statistical Association, 1617-1628.

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*. Vital and Health Statistics, Series 2, No. 158. Hyattsville, MD: National Center for Health Statistics.

Luiten, A., and Schouten, B. (2013). Tailored fieldwork design to increase representative household survey response: An experiment in the Survey of Consumer Satisfaction. *Journal of the Royal Statistical Society A*, 176, 169-189.

Lundquist, P., and Särndal, C.-E. (2013). Aspects of responsive design with applications to the Swedish Living Conditions Survey. *Journal of Official Statistics*, 29, 557-582.

Mohadjer, L., and Edwards, B. (2018). Paradata and dashboards in PIAAC. *Quality Assurance in Education*, 26, 263-277.

Peytchev, A., Baxter, R.K., and Carley-Baxter, L.R. (2009). Not all survey effort is equal: Reduction of nonresponse bias and nonresponse error. *Public Opinion Quarterly*, 73, 785-806.

Peytchev, A., Riley, S., Rosen, J., Murphy, J. and Lindblad, M. (2010). Reduction of nonresponse bias in surveys through case prioritization. *Survey Research Methods*, 4, 21-29.

Särndal, C.-E. (2011). The 2010 Morris Hansen Lecture: Dealing with survey nonresponse in data collection, in estimation. *Journal of Official Statistics*, 27, 1-21.

Särndal, C.-E., and Lundquist, P. (2014a). Balancing the response and adjusting estimates for nonresponse bias: Complementary activities. *Journal de la Société de Statistique*, 155, 28-50.

Särndal, C.-E., and Lundquist, P. (2014b). Accuracy in estimation with nonresponse: A function of the degree of imbalance and degree of explanation. *Journal of Survey Statistics and Methodology*, 2, 361-387.

Schaeffer, N.C., Dykema, J. and Maynard, D.W. (2010). Interviews and interviewing. In *Handbook of Survey Research*, Bingley, UK, Emerald Group Publishing, 437-470.

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/en/pub/12-001-x/2009001/article/10887-eng.pdf.

Schouten, B., Peytchev, A. and Wagner, J. (2017). *Adaptive Survey Design*, Boca Raton, FL: CRC Press.

Steeh, C., Kirgis, N., Cannon, B. and DeWitt, J. (2001). Are they really as bad as they seem? Nonresponse rates at the end of the twentieth century. *Journal of Official Statistics*, 17, 227-247.

Sturgis, P., Williams, J., Brunton-Smith, I. and Moore, J. (2017). Fieldwork effort, response rate, and the distribution of survey outcomes: A multilevel meta-analysis. *Public Opinion Quarterly*, 81, 523-542.

Tourangeau, R. (2017). Presidential address: Paradoxes of nonresponse. *Public Opinion Quarterly*, 81, 803-814.

Tourangeau, R., Kreuter, F. and Eckman, S. (2012). Motivated underreporting in screening surveys. *Public Opinion Quarterly*, 76, 453-469.

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 A*, 180, 203-223.

Tourangeau, R., Yan, T., Sun, H., Hyland, A. and Stanton, C.A. (2019). Population Assessment of Tobacco and Health (PATH) reliability and validity study: Selected reliability and validity estimates. *Tobacco Control*, 28, 663-668.

U.S. Census Bureau (2014). *American Community Survey Design and Methodology*. Downloaded February 19, 2016 from https://www.census.gov/programs-surveys/acs/methodology/design-and-methodology.html.

Wagner, J. (2013). Adaptive contact strategies in telephone and face-to-face surveys. *Survey Research Methods*, 7, 45-55.

Wagner, J., and Olson, K. (2018). An analysis of interviewer travel and field outcomes in two field surveys. *Journal of Official Statistics*, 34, 211-237.

Wagner, J., West, B.T., Kirgis, N., Lepkowski, J.M., Axinn, W.G. and Kruger Ndiaye, S. (2012). Use of paradata in a responsive design framework to manage a field data collection. *Journal of Official Statistics*, 28, 477-499.

Walejko, G., and Miller, P. (2015). The 2013 census test: Piloting methods to reduce 2020 Census costs. *Survey Practice*, 8.

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 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.

West, B.T., Conrad, F.G., Kreuter, F. and Mittereder, F. (2018). Nonresponse and measurement error variance among interviewers in standardized and conversational interviewing. *Journal of Survey Statistics and Methodology*, 6, 335-359.

Williams, D., and Brick, J.M. (2018). Trends in U.S. face-to-face household survey nonresponse and level of effort. *Journal of Survey Statistics and Methodology*, 6, 186-211.

Survey Methodology, June 2021
Vol. 47, No. 1, pp. 29-58
Statistics Canada, Catalogue No. 12-001-X

Integration of data from probability surveys and big found data for finite population inference using mass imputation

Shu Yang, Jae Kwang Kim and Youngdeok Hwang[1]

Abstract

Multiple data sources are becoming increasingly available for statistical analyses in the era of big data. As an important example in finite-population inference, we consider an imputation approach to combining data from a probability survey and big found data. We focus on the case when the study variable is observed in the big data only, but the other auxiliary variables are commonly observed in both data. Unlike the usual imputation for missing data analysis, we create imputed values for all units in the probability sample. Such mass imputation is attractive in the context of survey data integration (Kim and Rao, 2012). We extend mass imputation as a tool for data integration of survey data and big non-survey data. The mass imputation methods and their statistical properties are presented. The matching estimator of Rivers (2007) is also covered as a special case. Variance estimation with mass-imputed data is discussed. The simulation results demonstrate the proposed estimators outperform existing competitors in terms of robustness and efficiency.

Key Words: Calibration weighting; Data fusion; Generalized additive model; Matching; Nearest neighbor imputation; Post stratification.

1. Introduction

In finite population inference, probability sampling is the gold standard for obtaining a representative sample from the target population. Because the selection probability is known, the subsequent inference from a probability sample is often design-based and respect the way in which the data were collected; see Särndal, Swensson and Wretman (2003), Cochran (2007), Fuller (2009) for textbook discussions. However, large-scale survey programs continually face heightened demands coupled with reduced resources. Demands include requests for estimates for domains with small sample sizes and desires for more timely estimates. Simultaneously, program budget cuts force reductions in sample sizes, and decreasing response rates make nonresponse bias an important concern. Baker, Brick, Bates, Battaglia, Couper, Dever, Gile and Tourangeau (2013) address the current challenges in using probability samples for finite population inferences.

To meet the new challenges, statistical offices face the increasing pressure to utilize convenient but often uncontrolled big data sources (also called big found data), such as satellite information (McRoberts, Tomppo and Næsset, 2010), mobile sensor data (Palmer, Espenshade, Bartumeus, Chung, Ozgencil and Li, 2013), and web survey panels (Tourangeau, Conrad and Couper, 2013). Couper (2013), Citro (2014), Tam and Clarke (2015), and Pfeffermann, Eltinge and Brown (2015) articulate the promise of harnessing big data for official and survey statistics but also raise many issues regarding big data sources. While such data sources provide timely data for a large number of variables and population elements, they are non-probability samples and often fail to represent the target population of interest because of inherent selection biases. Tam and Kim (2018) also cover some ethical challenges of big data for official

1. Shu Yang, Department of Statistics, North Carolina State University, 2311 Stinson Drive, Campus Box 8203, Raleigh, NC 27695, U.S.A. E-mail: syang24@ncsu.edu; Jae Kwang Kim, Department of Statistics, 1208 Snedecor Hall, Iowa State University, Ames, IA 50011, U.S.A. Youngdeok Hwang, Paul H. Chook Department of Information Systems and Statistics, Baruch College, City University of New York, New York, NY 10010, U.S.A.

statisticians and discuss some preliminary methods of correcting for selection bias in big data. See Keiding and Louis (2016), Elliott and Valliant (2017), Buelens, Burger and van den Brakel (2018), and Beaumont (2020) for recent reviews of the challenges in using non-probability samples for inferences.

To utilize modern data sources in statistically defensible ways, it is important to develop statistical tools for data integration for combining a probability sample with big non-probability data. Data integration for finite population inference is similar to the problem of combining randomized clinical trial studies and non-randomized epidemiological studies for causal inference of treatment effects (Keiding and Louis, 2016). We are particularly interested in developing data integration under the setup where the study variable is observed in the big data only, but some other variables are commonly observed in both data. In this case, survey statisticians and biostatisticians have provided different methods for combining information from multiple data sources. Lohr and Raghunathan (2017), Yang and Kim (2020), and Rao (2020) provide a review of statistical methods of data integration for finite population inference. Existing methods for data integration can be categorized into three types as follows.

The first type is the so-called propensity score adjustment (Rosenbaum and Rubin, 1983). In this approach, the probability of a unit being selected into the big sample, which is referred to as the propensity score, is modeled and estimated for all units in the big data sample. The subsequent adjustments, such as propensity score weighting or stratification, can then be used to adjust for selection biases; see, e.g., Lee and Valliant (2009), Valliant and Dever (2011), Elliott and Valliant (2017). Stuart, Bradshaw and Leaf (2015), Stuart, Cole, Bradshaw and Leaf (2011), Buchanan, Hudgens, Cole, Mollan, Sax, Daar, Adimora, Eron and Mugavero (2018) use propensity score weighting to generalize results from randomized trials to a target population. O'Muircheartaigh and Hedges (2014) propose propensity score stratification for analyzing a nonrandomized social experiment. One of the notable disadvantages of the propensity score methods is that they rely on an explicit propensity score model and are biased if the model is mis-specified (Kang and Schafer, 2007).

The second type uses calibration (Deville and Särndal, 1992; Kott, 2006; Dong, Yang, Wang, Zeng and Cai, 2020). This technique can be used to calibrate auxiliary information in the big data sample with that in the probability sample, so that after calibration the big data sample is similar to the target population (DiSogra, Cobb, Chan and Dennis, 2011). Because calibration does not require parametric modeling, it is attractive to survey practitioners. However, this approach requires the information (such as the moments) of the auxiliary variables for the population is known or at least can be estimated from a probability sample.

The third type is mass imputation, where the imputed values are created for the whole elements in the probability sample. In the usual imputation for missing data analysis, the respondents in the sample provide a training dataset for developing an imputation model. In the mass imputation, an independent big data sample is used as a training dataset, and imputation is applied to all units in the probability sample. While the mass imputation idea for incorporating information from big data is very natural, the literature on mass imputation itself is sparse. Breidt, McVey and Fuller (1996) discuss mass imputation for two-

phase sampling. Rivers (2007) proposes a mass imputation approach using nearest neighbor imputation but the theory is not fully developed. Kim and Rao (2012) develop a rigorous theory for mass imputation using two independent probability samples. Chipperfield, Chessman and Lim (2012) discuss composite estimation when one of the surveys is mass imputed. Bethlehem (2016) discuss practical issues in sample matching. Recently, Kim and Wang (2019) develop a theory for mass imputation for big data using a parametric model approach. However, the parametric model assumptions do not necessarily hold in practice. In order for mass imputation to be more useful and practical, the assumptions should be as weak as possible.

We summarize our contributions in this paper below:

1. We first develop a formal framework for mass imputation incorporating information from big data into a probability sample and present rigorous asymptotic results for the mass imputation estimators. Our framework covers the nearest neighbor imputation estimator of Rivers (2007). Unlike Kim and Wang (2019), we do not make strong parametric model assumptions for mass imputation. Thus, the proposed method is appealing to survey practitioners.
2. We also investigate two strategies for improving the nearest neighbor imputation estimator, one using k nearest neighbor imputation (Mack and Rosenblatt, 1979) and the other using generalized additive models (Wood, 2006). In k nearest neighbor imputation, instead of using one nearest neighbor, we identify multiple nearest neighbors in the big data sample and use the average response as the imputed value. This method is popular in the international forest inventory community for combining ground-based observations with imagines from remote sensors (McRoberts et al., 2010). In this paper, we establish asymptotic results for the k nearest neighbor estimator. In the second strategy, we investigate modern techniques of prediction for mass imputation with flexible models. We use generalized additive models (Wood, 2006) to learn the relationship of the outcome and covariates from the big data and create predictions for the probability samples. We note that this strategy can apply to a wider class of semi- and non-parametric estimators such as single index models, Lasso estimators (Belloni, Chernozhukov, Chetverikov and Kato, 2015), and machine learning methods such as random forests (Breiman, 2001).
3. Using a novel calibration weighting idea, we propose an efficient mass imputation estimator and develop its asymptotic results. The efficiency gain is justified under a purely design-based framework and no model assumptions are used. We consider the case when additionally the membership to the big data can be determined throughout the probability sample. The key insight is that the subsample of units in Sample A with the big data membership constitutes a second-phase sample from the big data sample, which acts as a new population. We calibrate the information in the second-phase sample to be the same as the new acting population. The calibration process in turn improves the accuracy of the mass imputation estimator without specifying any model assumptions.

The structure of the paper is as follows. In Section 2, we introduce the basic setup. In Section 3, we present the methodology for the nearest neighbor imputation and establish its asymptotic properties. In Section 4, we investigate two strategies for improving the nearest neighbor imputation estimator, one using k nearest neighbor imputation and the other using generalized additive models. In Section 5, we propose a regression calibration technique to improve the efficiency of the mass imputation estimators when additionally the big data membership is observed throughout the probability sample. In Section 6, we demonstrate that the proposed estimators are robust and efficient by simulation studies based on artificial data and real-life data from U.S. Census Bureau's Monthly Retail Trade Survey. In Section 7, we present a case study applying the proposed method to integrate national health survey data and national health insurance records. Section 8 concludes with a discussion.

2. Basic setup

2.1 Notation: Two data sources

Let $\mathcal{F}_N = \{(\mathbf{X}_i, Y_i): i \in U\}$ with $U = \{1, \ldots, N\}$ denote a finite population, where $\mathbf{X}_i = (X_i^1, \ldots, X_i^p)$ is a p-dimensional vector of covariates, and Y_i is the study variable. We assume that $\mathcal{F}_N$ is a random sample from a superpopulation model ζ, and N is known. Our objective is to estimate the general finite population parameter $\mu_g = N^{-1}\sum_{i=1}^N g(Y_i)$ for some known $g(\cdot)$. For example, if $g(Y) = Y$, $\mu_g = N^{-1}\sum_{i=1}^N Y_i$ is the population mean of Y. If $g(Y) = \mathbf{1}(Y < c)$ for some constant c, $\mu_g = N^{-1}\sum_{i=1}^N \mathbf{1}(Y_i < c)$ is the population proportion of Y less than c.

Suppose that there are two data sources, one from a probability sample, referred to as Sample A, and the other from a big data source, referred to as Sample B. Table 2.1 illustrates the observed data structure. Sample A contains observations $\mathcal{O}_A = \{(d_i = \pi_i^{-1}, \mathbf{X}_i): i \in A\}$ with sample size $n = |A|$, where $\pi_i = P(i \in A)$ is known throughout Sample A, and Sample B contains observations $\mathcal{O}_B = \{(\mathbf{X}_i, Y_i): i \in B\}$ with sample size $N_B = |B|$. Often the probability sample contains many other items but we only use those items overlapping with our big data. Although the big data source has a large sample size, the sampling mechanism is often unknown, and we cannot compute the first-order inclusion probability for Horvitz-Thompson estimation. The naive estimators without adjusting for the sampling process are subject to selection biases. On the other hand, although the probability sample with sampling weights represents the finite population, it does not observe the study variable.

Table 2.1
Two data sources. "√" and "?" indicate observed and unobserved data, respectively

		Sample weight $d = \pi^{-1}$	Covariate $\mathbf{X}$	Study Variable Y
Probability Sample	1	√	√	?
$\mathcal{O}_A$	⋮	⋮	⋮	⋮
	n	√	√	?
Big Data Sample	1	?	√	√
$\mathcal{O}_B$	⋮	⋮	⋮	⋮
	N_B	?	√	√

Sample A is a probability sample, and Sample B is a big data but may have selection biases.

2.2 Assumptions

Let $f(Y|\mathbf{X})$ be the conditional density function of Y given $\mathbf{X}$ in the superpopulation model ζ. Let $f(\mathbf{X})$ and $f(\mathbf{X}|\delta_B = 1)$ be the density function of $\mathbf{X}$ in the finite population and Sample B, respectively, where δ_B is the indicator of selection to Sample B. We first make the following assumptions.

Assumption 1 (Ignorability). *Conditional on* $\mathbf{X}$, *the density of* Y *in Sample B follows the superpopulation model; i.e.,* $f(Y|\mathbf{X}; \delta_B = 1) = f(Y|\mathbf{X})$.

Assumptions 1 and 2 constitute the strong ignorability condition (Rosenbaum and Rubin, 1983). This setup has previously been used by several authors; see, e.g., Rivers (2007), Vavreck and Rivers (2008). Assumption 1 states the ignorability of the selection mechanism to Sample B conditional upon the covariates. Assumption 1 also implies that $P(\delta_B = 1|\mathbf{X}, Y) = P(\delta_B = 1|\mathbf{X})$. This assumption holds if the set of covariates contains all predictors for the outcome that affect the possibility of being selected in Sample B. Under this assumption, the missing outcomes in Sample A are missing at random (Rubin, 1976).

Assumption 2 (Common support). *The vector of covariates* $\mathbf{X} \in \mathbb{R}^p$ *has a compact and convex support, with its density bounded and bounded away from zero. There exist constants* C_l *and* C_u *such that* $C_l \leq f(\mathbf{X})/f(\mathbf{X}|\delta_B = 1) \leq C_u$ *almost surely.*

Assumption 2 implies that the support of $\mathbf{X}$ in Sample B is the same as that in the finite population. This assumption can also be formulated as a positivity assumption that $P(\delta_B = 1|\mathbf{X}) > 0$ for all $\mathbf{X}$. Assumption 2 does not hold if certain units would never be included in the big data sample. The plausibility of this assumption can be judged by subject matter knowledge. For diagnosis purpose, we can examine the distribution of the estimated propensity scores or the distribution of the propensity score weights in Sample A. Values of propensity score close to zero or extreme large values of the propensity score weights indicate the possible positivity violation. We assume all covariates are continuous. Categorical variables can be handled by first defining imputation classes using the partition of the categories and then estimating the average of the outcome using the nearest neighbor imputation within imputation classes. In our context, Sample B is a big data sample and therefore the size of donors for each imputation class can be reasonable large.

3. Methodology

3.1 Nearest neighbor imputation

For simplicity, we will focus on the Horvitz-Thompson type estimator, although our discussion applies to other type of estimators. If Y_i were observed throughout Sample A, the Horvitz–Thompson estimator

$\hat{\mu}_{g,\mathrm{HT}} = N^{-1}\sum_{i\in A}\pi_i^{-1}g(Y_i)$ can be used. We consider the imputation estimator of μ_g, given by $\hat{\mu}_{g,I} = N^{-1}\sum_{i\in A}\pi_i^{-1}g(Y_i^*)$, where Y_i^* is an imputed value for Y_i. Creating imputed values for the whole data is called mass imputation (Chipperfield et al., 2012; Kim and Rao, 2012).

To find suitable imputed values, we consider nearest neighbor imputation; that is, find the closest matching unit from Sample B based on the $\mathbf{X}$ values and use the corresponding Y value from this unit as the imputed value. This approach has been called *Sample Matching* by Rivers (2007). To investigate the theoretical properties, we first consider matching with replacement with single imputation; the discussion on k nearest neighbor imputation is presented in Section 4.

The nearest neighbor approach to mass imputation can be described in the following steps:

Step 1. For each unit $i \in A$, find the nearest neighbor from Sample B with the minimum distance between $\mathbf{X}_j$ and $\mathbf{X}_i$. Let $i(1)$ be the index of its nearest neighbor, which satisfies $d(\mathbf{X}_{i(1)}, \mathbf{X}_i) \le d(\mathbf{X}_j, \mathbf{X}_i)$, for $j \in B$, where $d(\mathbf{X}_i, \mathbf{X}_j)$ is a distance function between $\mathbf{X}_i$ and $\mathbf{X}_j$. If there are ties, randomly select one as the nearest neighbor. Without loss of generality, we use the Euclidean distance, $d(\mathbf{X}_i, \mathbf{X}_j) = \|\mathbf{X}_i - \mathbf{X}_j\|$, where $\|\mathbf{X}\| = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{1/2}$, to determine neighbors.

Step 2. The nearest neighbor imputation estimator of μ_g is

$$\hat{\mu}_{g,\,\mathrm{nni}} = \frac{1}{N}\sum_{i\in A}\pi_i^{-1}g(Y_{i(1)}). \tag{3.1}$$

Remark 1. *Our theoretical development applies to a general class of distances* $\|\mathbf{X}\|_\Sigma = (\mathbf{X}^{\mathrm{T}}\Sigma^{-1}\mathbf{X})^{1/2}$, *where* Σ *is a positive definite matrix (Abadie and Imbens, 2006). This class includes the standard Mahalanobis distance by taking* Σ *to be the empirical covariance matrix of* $\mathbf{X}$. *Write* $\Sigma = L^{\mathrm{T}}L$. *Notice that* $\|\mathbf{X}\|_\Sigma = \{(L\mathbf{X})\,L\mathbf{X}\}^{1/2} = \|L\mathbf{X}\|$. *Hence, using* $\|\cdot\|_\Sigma$ *and* $\mathbf{X}$ *is equivalent to using* $\|\cdot\|$ *and* $L\mathbf{X}$. *So, we can carry over the the theoretical result to the case with* $\|\mathbf{X}\|_\Sigma$.

Comparing to model-based imputation, nearest neighbor imputation has several advantages. First, it does not require strong parametric model assumptions and therefore is robust to model misspecification. Second, nearest neighbor imputation is donor-based, where the imputed value is a value that was actually measured and will always be within the bounds of observed values. Third, in contrast to regression imputation approaches, nearest neighbor imputation can retain the complex variance covariance structure of the data. Moreover, for the same imputed dataset, one can estimate different parameters by choosing reasonable $g(\cdot)$. Recall that p is the dimension of $\mathbf{X}$. The asymptotic bias of $\hat{\mu}_{g,\mathrm{nni}}$ is of order $O_{\mathrm{p}}(N_B^{-1/p})$ (Abadie and Imbens, 2006), which is negligible when the number of continuous covariates is fixed at a reasonable number and the size of the matching donor pool is huge as in our big data setup. In the presence of a large dimension of $\mathbf{X}$, variable selection is necessary for the nearest neighbor imputation estimator to have good statistical properties. In this case, we suggest selecting important variables that are associated with the outcome in order to ensure Assumption 1 holds and also to increase estimation precision (Brookhart, Schneeweiss, Rothman, Glynn, Avorn and Stürmer, 2006).

3.2 Asymptotic results

To study the asymptotic properties of $\hat{\mu}_{g,\text{nni}}$, we impose the following regularity conditions.

Assumption 3. *(i)* $f(\mathbf{X})$ *and* $\mu_g(\mathbf{X}) = E\{g(Y)|\mathbf{X}\}$ *are continuously differentiable for any continuous and bounded* $g(Y)$, *and (ii)* $E\{g(Y)^\beta|\mathbf{X}\}$ *is bounded for* $\beta = 1, 2$.

Assumption 4. *(i) There exist positive constants* C_1 *and* C_2 *such that* $C_1 \le Nn^{-1}\pi_i \le C_2$, *for* $i = 1, \ldots, N$; *(ii) the sampling fraction for Sample A is negligible,* $nN^{-1} = o(1)$; *and (iii) the sequence of the Horvitz-Thompson estimators* $\hat{\mu}_{g,\text{HT}}$ *satisfies* $\text{var}_p(\hat{\mu}_{g,\text{HT}}) = O(n^{-1})$ *and* $\{\text{var}_p(\hat{\mu}_{g,\text{HT}})\}^{-1/2}(\hat{\mu}_{g,\text{HT}} - \mu_g)|\mathcal{F}_N \to \mathcal{N}(0,1)$ *in distribution, as* $n \to \infty$, *where* $\text{var}_p(\cdot) = \text{var}(\cdot|\mathcal{F}_N)$ *is the variance under the sampling design for Sample A.*

For clarification, the probability distribution underpinning the notation $E(\cdot)$, $\text{var}(\cdot)$, $o_\text{p}(\cdot)$ and $O_\text{p}(\cdot)$ is the joint distribution of the superpopulation model and the sampling processes for Samples A and B. Assumption 3 is a technical condition imposed on the functional continuity and finite moments, which holds for common models; see, e.g., Mack (1981). Assumption 4 holds for standard sampling designs in survey practice (Fuller, 2009; Chapter 1). It requires the sampling weights to behave well in the sense that there do not exist extremely large weights that dominate other weights. This occurs when subjects when certain characteristics are largely underrepresented in the sample. Sufficient conditions for Assumption 4 (iii) can be found in Chapter 3 of Fuller (2009).

We derive the asymptotic theory for $\hat{\mu}_{g,\text{nni}}$ in the following theorem and defer its proof to the Supplementary Material.

Theorem 1. *Under Assumptions 1–3 and* $NN_B^{-1} = O(1)$, $\hat{\mu}_{g,\text{nni}}$ *has the same distribution as* $\hat{\mu}_{g,\text{HT}}$ *as* $N_B \to \infty$. *Furthermore, under Assumption 4,* $\hat{\mu}_{g,\text{nni}}$ *is consistent for* μ_g, *and*

$$n^{1/2}(\hat{\mu}_{g,\text{nni}} - \mu_g) \to \mathcal{N}(0, V_\text{nni}), \tag{3.2}$$

where

$$V_\text{nni} = \lim_{n\to\infty} \frac{n}{N^2} E\left[\text{var}_p\left\{\sum_{i\in A} \pi_i^{-1} g(Y_i)\right\}\right].$$

Theorem 1 implies that the standard point estimator can be applied to the imputed data $\{(\mathbf{X}_i, Y_{i(1)}): i \in A\}$ as if the $Y_{i(1)}$'s were observed values. Let π_{ij} be the joint inclusion probability for units i and j. We show in the Supplementary Material that the direct variable estimator based on the imputed data

$$\hat{V}_\text{nni} = \frac{n}{N^2} \sum_{i\in A}\sum_{j\in A} \frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j} \frac{g(Y_{i(1)})}{\pi_i} \frac{g(Y_{j(1)})}{\pi_j}$$

is consistent for V_nni.

4. Other techniques for mass imputation

4.1 k -nearest neighbor imputation

Instead of using a single imputed value, we now consider fractional imputation with k imputed values for each missing outcome. Fractional imputation is designed to reduce the variance of the final estimator due to imputation (Kalton and Kish, 1984; Kim and Fuller, 2004).

Assume no matching ties, let $\mathcal{J}_k(i)$ be the set of k nearest neighbors for unit i

$$\mathcal{J}_k(i) = \left\{ l \in B : \sum_{j \in B} 1_{\{d(\mathbf{x}_j, \mathbf{x}_i) \le d(\mathbf{x}_l, \mathbf{x}_i)\}} \le k \right\} = \{ i(1), \ldots, i(k) \}.$$

The k nearest neighbor approach to mass imputation can be described in the following steps:

Step 1. For each unit $i \in A$, find the k nearest neighbors from Sample B, $\mathcal{J}_k(i)$. Impute the Y value for unit i by $\hat{\mu}_g(\mathbf{X}_i) = k^{-1} \sum_{j=1}^{k} g(Y_{i(j)})$.

Step 2. The k nearest neighbor imputation estimator of μ_g is

$$\hat{\mu}_{g,\text{knn}} = \frac{1}{N} \sum_{i \in A} \pi_i^{-1} \hat{\mu}_g(\mathbf{X}_i). \tag{4.1}$$

In the non-parametric estimation literature, researchers have investigated the asymptotic properties of the k nearest neighbor imputation estimators extensively. See, e.g., Mack and Rosenblatt (1979) and Mack (1981) for early references. Cheng (1994) establishes root-n consistency of the k nearest neighbor imputation estimator of the outcome mean when the outcome is subject to missingness. We derive the asymptotic theory for $\hat{\mu}_{g,\text{knn}}$ in the context of mass imputation combining a probability sample and a big data sample in the following theorem and defer its proof to the Supplementary Material.

Theorem 2. *Under Assumptions 1–4,* $n(k/N)^{4/p} \to 0$, $k/n \to 0$, *and* $k^2/n \to \infty$,

$$n^{1/2}(\hat{\mu}_{g,\text{knn}} - \mu_g) \to \mathcal{N}(0, V_{\text{knn}}), \tag{4.2}$$

where

$$V_{\text{knn}} = \lim_{n \to \infty} \frac{n}{N^2} \left(E\left[\text{var}_p \left\{ \sum_{i \in A} \pi_i^{-1} \mu_g(\mathbf{X}_i) \right\} \right] + E\left\{ \frac{1 - \pi_B(\mathbf{X})}{\pi_B(\mathbf{X})} \sigma_g^2(\mathbf{X}) \right\} \right),$$

and $\pi_B(\mathbf{X}) = P(\delta_B = 1 | \mathbf{X})$ *and* $\sigma_g^2(\mathbf{X}) = \text{var}\{g(Y) | \mathbf{X}\}$.

If $\pi_B(\mathbf{X})$ goes to 1, V_{knn} reduces to $\lim_{n \to \infty} (n/N^2) E\left[\text{var}_p \left\{ \sum_{i \in A} \pi_i^{-1} \mu_g(\mathbf{X}_i) \right\} \right]$. It suggests that if the big sample is a large fraction of the target population, V_{knn} can be smaller than V_{nni}, suggesting that $\hat{\mu}_{g,\text{knn}}$ gains efficiency over $\hat{\mu}_{g,\text{nni}}$. In finite samples, Beretta and Santaniello (2016) conduct a simulation study to compare nearest neighbor imputation and k nearest neighbor imputation in the setting with independent and identically distributed data. They found that k nearest neighbor imputation with a small k outperforms nearest neighbor imputation in terms of mean squared error. On the one hand, a larger k can use more information in the big data sample and leads to more efficiency gain; on the other hand, k

cannot be too large, in order to control the bias of our estimator. In practice, we suggest using data-driven methods, such as cross-validation, to choose a reasonable k, and conducting sensitivity analysis varying the choice of k.

4.2 Generalized additive models

Nearest neighbor imputation methods are non-parametric. On the other hand, parametric models especially linear models are sensitive to model misspecification. We now consider semiparametric methods for mass imputation. Among semiparametric methods, generalized additive models (Hastie and Tibshirani, 1990) are flexible regarding model specification of the dependence of Y on $\mathbf{X}$ by specifying the model only through smooth functions rather than assuming a parametric relationship. As other non-parametric methods, the performance of generalized additive models will deteriorate as the dimension of $\mathbf{X}$ becomes large. For $\mathbf{X}$ with a moderate dimension, we apply generalized additive models to leverage the predictive power of the big data sample to produce a predictive model for Y given $\mathbf{X}$, so as to facilitate mass imputation for the probability sample.

We assume that $g(Y_i)$ given $\mathbf{X}_i$ follows some exponential family distribution, and

$$h^{-1}\{\mu_g(\mathbf{X}_i)\} = f_1(X_i^1) + f_2(X_i^2) + \ldots f_p(X_i^p), \tag{4.3}$$

where $h(\cdot)$ is an inverse link function, and each $f_k(\cdot)$ is a smooth function of X^k, for $k = 1, \ldots, p$. Model (4.3) allows for rather flexible specification of the dependence of Y on $\mathbf{X}$. The estimated function $f_k(X^k)$ can reveal possible nonlinearities of the relationship of Y and X^k.

There are several challenges in fitting model (4.3). First, $f_k(x)$ is an infinite-dimensional parameter, estimation of which often relies on some approximation. Second, we need to decide how smooth the $f_k(x)$ should be to balance the trade-off between model complexity and overfitting to the data at hand.

To solve the first issue, a common way to approximate $f_k(x)$ using splines. Let $B_m(x)$ be the basis spline functions for $m = 1, \ldots, M$ (Ruppert, Wand and Carroll, 2009). We approximate $f_k(x)$ by $f_k(x) = \sum_{m=1}^{M} \gamma_m^k B_m(x)$ with spline coefficients γ_m^k. This leads to an approximation of model (4.3):

$$h^{-1}\left[\hat{E}\{g(Y_i)\,|\,\mathbf{X}_i\}\right] = \sum_{k=1}^{p}\sum_{m=1}^{M} \gamma_m^k B_m(X_i^k). \tag{4.4}$$

In (4.4), a large M allows for increased model complexity and also an increased chance of overfitting; while a small M may result in an inadequate model. This trade-off is balanced by choosing a relatively large M and then penalizing the model complexity in the estimation stage (Eilers and Marx, 1996). Let the vector of spline coefficients be $\gamma_k^{\mathrm{T}} = (\gamma_1^k, \ldots, \gamma_m^k)$ and $\gamma^{\mathrm{T}} = (\gamma_1^{\mathrm{T}}, \ldots, \gamma_p^{\mathrm{T}})$. The estimate $\hat{\gamma}$ is obtained by maximizing the penalized likelihood:

$$-2l(\gamma) + \sum_{k=1}^{p} \lambda_k \gamma_k^{\mathrm{T}} S_k \gamma_k \tag{4.5}$$

where $l(\gamma)$ is the log likelihood function of γ, S_k is a matrix with the $(m, l)^{\text{th}}$ component $\int B_m''(x) B_l''(x)\,\mathrm{d}x$, $\gamma_k^{\mathrm{T}} S_k \gamma_k$ regularizes f_k to be smooth for which the degree of smoothness is controlled

by λ_k. Given the smoothing parameter $\lambda^{\mathrm{T}} = (\lambda_1, \ldots, \lambda_p)$, the penalized likelihood function in (4.5) is optimized by a penalized version of the iteratively reweighted least squares algorithm (Nelder and Baker, 1972; McCullagh, 1984) to obtain $\hat{\gamma}$. Regarding the choice of λ, we note that λ controls the trade-off between model complexity and overfitting, which can be estimated separately from other model coefficients using generalized cross-validation or estimated simultaneously using restricted maximum likelihood estimation (Wood, 2006). In practice, the model performance is not sensitive to the choice of the number of basis functions as long as the number of basis functions is large relative to the sample size in the specification, but rather estimation of the smoothing parameter is critical to control the model complexity.

Once fitting the model, we can create an imputed value for each element i in Sample A as

$$\hat{\mu}_{g,\mathrm{GAM}}(\mathbf{X}_i) = h\left\{\hat{f}_1(X_i^1) + \hat{f}_2(X_i^2) + \ldots \hat{f}_p(X_i^p)\right\},$$

where $\hat{f}_k(x) = \sum_{m=1}^{M} \hat{\gamma}_m^k B_m(x)$ for $k = 1, \ldots, p$. The mass imputation estimator based on the generalized additive model is

$$\hat{\mu}_{g,\,\mathrm{GAM}} = \frac{1}{N}\sum_{i\in A} \pi_i^{-1} \hat{\mu}_{g,\mathrm{GAM}}(\mathbf{X}_i).$$

Because in our context, the sample size of Sample B is much larger than that of Sample A, the estimation error in the imputation model can be negligible compared to the sampling variability of $\hat{\mu}_{g,\mathrm{GAM}}$.

To close this subsection, it is worth commenting on the assumption of additive effects of $\mathbf{X}$ in model (4.3). This assumption may be fairly strong one. To relax the additivity assumption, we can extend model (4.3) to include interactions through using the tensor product basis. For example, we can include a bivariate interaction surface $f_{12}(X^1, X^2) = \sum_{m=1}^{M}\sum_{l=1}^{L} \gamma_{ml} B_m(X^1) B_l(X^2)$. When using the tensor product basis, care should be taken with respect to the penalty function in order to result in appropriate effective degrees of freedom for the smoother. This topic has been investigated extensively in the literature; see, e.g., Wood (2006).

5. Regression calibration

In practice, especially for government agencies, one nearest neighbor may be preferred because of its simplicity in implementation and data storage. We now consider another strategy to improve the efficiency for $\hat{\mu}_{g,\mathrm{nni}}$ when additionally the membership to Sample B can be determined throughout Sample A with the indicator δ_B. In some situation, we can obtain δ_B by matching the membership to Sample B (i.e., data linkage). We focus on the ideal setting without linkage errors. The key insight is that the subsample of units in Sample A with $\delta_B = 1$ constitutes a second-phase sample from Sample B, where Sample B acts as a new population. Standard regression calibration requires all calibration variables to be observed in Sample A and Sample B, and thus rules out the possibility of using Y as the calibration variable due to lack of the outcome data from Sample B. One of the advantages of mass imputations is that we can leverage the imputed outcomes to facilitate calibration of Y.

Let $\mathbf{h}(\delta_B, \mathbf{X}, Y)$ be a multi-dimensional function of δ_B, $\delta_B\mathbf{X}$ and $\delta_B Y$, e.g., $\mathbf{h}(\delta_B, \mathbf{X}, Y) = (\delta_B, 1-\delta_B, \delta_B\mathbf{X}, \delta_B Y)^{\mathrm{T}}$. For simplicity of notation, we use $\mathbf{h}_i$ to denote $\mathbf{h}(\delta_{Bi}, \mathbf{X}_i, Y_i)$ and $\mathbf{h}_i^*$ to denote $\mathbf{h}(\delta_{Bi}, \mathbf{X}_i, Y_{i(1)})$. We can calculate the population quantity $\mathbf{H} = N^{-1}\sum_{i=1}^{N}\mathbf{h}_i$ from Sample B. This insight enables the typical calibration weighting in survey sampling with known marginal totals. In Sample A, we treat the imputed values as observed values, and the design weighted estimator of $\mathbf{H}$ is $\hat{\mathbf{H}}_A = N^{-1}\sum_{i\in A}\pi_i^{-1}\mathbf{h}_i^*$. In general, $\hat{\mathbf{H}}_A$ is not equal to $\mathbf{H}$. We can use the known information $\mathbf{H}$ to improve the efficiency of $\hat{\mu}_{g,\mathrm{nni}}$.

This suggests the following calibration strategy. We modify the original design weights $\{d_i: i \in A\}$ in $\hat{\mu}_{g,\mathrm{nni}}$ to a new set of weights $\{\omega_i: i \in A\}$ by minimizing a distance function

$$\sum_{i\in A} G(\omega_i, d_i) = \sum_{i\in A} d_i\left(\frac{\omega_i}{d_i} - 1\right)^2, \tag{5.1}$$

subject to the calibration constraints $N^{-1}\sum_{i\in A}\omega_i\mathbf{h}_i^* = \mathbf{H}$. By Lagrange multiplier, the solution to the constraint minimization problem is

$$\omega_i = d_i + \left(N\times\mathbf{H} - \sum_{k\in A} d_k\mathbf{h}_k^*\right)^{\mathrm{T}}\left(\sum_{k\in A} d_k\mathbf{h}_k^*\mathbf{h}_k^{*\mathrm{T}}\right)^{-1} d_i\mathbf{h}_i^*,$$

for $i \in A$. The resulting weights $\{\omega_i: i \in A\}$ can be called generalized regression weights.

The proposed estimator utilizing the new set of weights is

$$\hat{\mu}_{g,\mathrm{RC}} = \frac{1}{N}\sum_{i\in A}\omega_i g(Y_{i(1)}), \tag{5.2}$$

which is asymptotically equivalent to a generalized regression estimator (Park and Fuller, 2012). Following Yang and Ding (2020), one can show that $\hat{\mu}_{g,\mathrm{RC}}$ is the optimal estimator among the class of $\left\{\hat{\mu}_{g,\mathrm{nni}} + \left(N\times\mathbf{H} - \sum_{k\in A} d_k\mathbf{h}_k^*\right)^{\mathrm{T}}\boldsymbol{\gamma}: \boldsymbol{\gamma}\in\mathbb{R}^{\dim(\mathbf{h})}\right\}$.

We derive the asymptotic theory for $\hat{\mu}_{g,\mathrm{RC}}$ in the following theorem and defer its proof to the Supplementary Material.

Theorem 3. *Under Assumptions 1-4,*

$$n^{1/2}(\hat{\mu}_{g,\mathrm{RC}} - \mu_g) \to \mathcal{N}(0, V_{\mathrm{RC}}), \tag{5.3}$$

in distribution, as $n \to \infty$, *where*

$$V_{\mathrm{RC}} = \lim_{n\to\infty}\frac{n}{N^2} E\left(\mathrm{var}_p\left[\sum_{i\in A}\pi_i^{-1}\{g(Y_i) - \mathbf{h}_i^{\mathrm{T}}\boldsymbol{\beta}_N\}\right]\right),$$

and $\boldsymbol{\beta}_N = \left(\sum_{i=1}^{N}\mathbf{h}_i\mathbf{h}_i^{\mathrm{T}}\right)^{-1}\sum_{i=1}^{N}\mathbf{h}_i g(Y_i)$.

The calibrated estimator $\hat{\mu}_{g,\mathrm{RC}}$ improves the efficiency of $\hat{\mu}_{g,\mathrm{nni}}$ in the sense that V_{RC} is at most as large as V_{nni} given in Theorem 1. If $\mathbf{h}_i$ explains a proportion of the variability of $g(Y_i)$, V_{RC} is strictly less than V_{nni} and the efficiency gain does not require any parametric model assumption.

Remark 2 (Choice of distance functions). *Different distance functions in (5.1) can be considered. If we choose* $G(\omega_i, d_i) = -d_i\log(\omega_i/d_i)$, *it leads to empirical likelihood estimation (Newey and Smith,*

2004). If we choose the Kullback-Leibler distance function $G(\omega_i, d_i) = \omega_i \log(d_i/\omega_i)$, *it leads to exponential tilting estimation (Kitamura and Stutzer, 1997; Imbens, Johnson and Spady, 1998; Schennach, 2007; Dong et al., 2020). Under mild conditions, these procedures provide a set of weights that is asymptotically equivalent to the set of regression weights (Deville and Särndal, 1992; Breidt and Opsomer, 2017).*

For variance estimation, by Theorem (3), we construct a consistent variance estimator for $\hat{\mu}_{g,\mathrm{RC}}$ as $\hat{V}_{\mathrm{RC}}/n$, where

$$\hat{V}_{\mathrm{RC}} = \frac{n}{N^2}\sum_{i\in A}\sum_{j\in A}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\frac{\hat{e}_i}{\pi_i}\frac{\hat{e}_j}{\pi_j},$$

with $\hat{e}_i = g\left(Y_{i(1)}\right) - \mathbf{h}_i^{*\mathrm{T}}\hat{\boldsymbol{\beta}}$, and

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N}\mathbf{h}_i^{*}\mathbf{h}_i^{*\mathrm{T}}\right)^{-1}\begin{pmatrix}\sum_{i=1}^{N}\delta_{Bi}g(Y_i)\\ \sum_{i\in A}\pi_i^{-1}(1-\delta_{Bi})\,g\left(Y_{i(1)}\right)\\ \sum_{i=1}^{N}\delta_{Bi}\mathbf{X}_i g(Y_i)\\ \sum_{i=1}^{N}\delta_{Bi}Y_i g(Y_i)\end{pmatrix}.$$

6. Empirical experiments

In this section, we evaluate the finite sample performance of the proposed estimator using simulation studies, one based on artificial data using simple random sampling and the other based on a synthetic population file from a single month sample of the U.S. Census Bureau's Monthly Retail Trade Survey using stratified sampling.

6.1 Kim-Wang example

We use the simulation example in Kim and Wang (2019) to compare various estimators. We generate the data according to the following mechanism. We first generate a finite population $\mathcal{F}_N = \{\mathbf{X}_i = (X_{1i}, X_{2i}), \mathbf{Y}_i = (Y_{1i}, Y_{2i}): i = 1, \ldots, N\}$ with size N = 1,000,000, where Y_{1i} is a continuous outcome and Y_{2i} is a binary outcome. From the finite population, we select a big data Sample B where the inclusion indicator $\delta_{Bi} \sim \mathrm{Ber}(p_i)$ with p_i the inclusion probability for unit i with the sample size around 700,000. We obtain a representative Sample A of size n = 1,000 using simple random sampling. The parameters of interest are the population mean $N^{-1}\sum_{i=1}^{N}\mathbf{Y}_i$ and the conditional population mean of Y_1 given $Y_2 = 1$.

For generating the finite population, we consider linear models

$$Y_{1i} = 1 + X_{1i} + X_{2i} + \alpha_i + \varepsilon_i, \tag{6.1}$$
$$P(Y_{2i} = 1 \mid X_{1i}, X_{2i}; \alpha_i) = \mathrm{logit}(1 + X_{1i} + X_{2i} + \alpha_i),$$

and nonlinear models

$$Y_{1i} = 0.5(X_{1i} - 1.5)^2 + X_{2i}^2 + \alpha_i + \varepsilon_i, \tag{6.2}$$

$$P(Y_{2i} = 1 \mid X_{1i}, X_{2i}; \alpha_i) = \text{logit}\left\{0.5(X_{1i} - 1.5)^2 + X_{2i}^2 + \alpha_i\right\},$$

where $X_{1i} \sim \mathcal{N}(1, 1)$, $X_{2i} \sim \text{Exp}(1)$, $\alpha_i \sim \mathcal{N}(0, 1)$, $\varepsilon_i \sim \mathcal{N}(0, 1)$, and X_{1i}, X_{2i}, α_i and ε_i are mutually independent. The variables α_i induce the dependence of Y_{1i} and Y_{2i} even adjusting for X_{1i} and X_{2i}. For the big-data inclusion probability, we also consider a logistic linear model

$$\text{logit}(p_i) = X_{2i}, \tag{6.3}$$

and a nonlinear logistic model

$$\text{logit}(p_i) = -3 + (X_{1i} - 1.5)^2 + (X_{2i} - 2)^2. \tag{6.4}$$

We consider the following combinations: I. (6.1) and (6.3); II. (6.1) and (6.4); II. (6.2) and (6.3); and IV. (6.2) and (6.4) for data generating mechanisms. Therefore, the simulation setup is a 2×2 factorial design with two levels in each factor.

Chen, Li and Wu (2020) propose the inverse propensity score weighting estimator using the estimated probability of selection into Sample B and the doubly robust estimator which further incorporates an outcome regression model. To evaluate the robustness and efficiency, we compare the following estimators:

1. $\hat{\mu}_{\text{HT}}$, the Horvitz–Thompson estimator assuming the Y_i's were observed in Sample A for the purpose of benchmark comparison;
2. $\hat{\mu}_{\text{ipw}}$, the inverse propensity score weighting estimator,

   $$\hat{\mu}_{\text{ipw}} = \frac{1}{N}\sum_{i \in B} \frac{1}{p_i(\hat{\eta})} Y_i,$$

   where $p_i(\eta) = P(\delta_{Bi} = 1 \mid X_{2i}; \eta)$ is a logistic regression model with the linear predictor X_{2i} with an unknown parameter η, and $\hat{\eta}$ is an estimator of η obtained by maximizing the modified likelihood function of η (Chen et al., 2019) based on Samples A and B;
3. $\hat{\mu}_{\text{dr}}$, the doubly robust estimator of Chen et al. (2019),

   $$\hat{\mu}_{\text{dr}} = \frac{1}{N}\sum_{i \in B} \frac{1}{p_i(\hat{\eta})}\left(Y_i - \mathbf{X}_i^{\text{T}}\hat{\boldsymbol{\beta}}\right) + \frac{1}{n}\sum_{i \in A} \mathbf{X}_i^{\text{T}}\hat{\boldsymbol{\beta}},$$

   where $\hat{\boldsymbol{\beta}}$ is the estimated regression coefficients using (6.1) as the working outcome regression model based on Sample B;
4. $\hat{\mu}_{\text{nni}}$, the nearest neighbor imputation estimator;
5. $\hat{\mu}_{\text{knn}}$, the k nearest neighbor imputation estimator with $k = 5$;
6. $\hat{\mu}_{\text{GAM}}$, the generalized additive model imputation estimator;
7. $\hat{\mu}_{\text{RC}}$, the regression calibration estimator based on $\hat{\mu}_{\text{nni}}$ with calibration variables $\mathbf{H}(\delta_B, \mathbf{X}, Y) = (\delta_B, 1 - \delta_B, \delta_B\mathbf{X}, \delta_B Y)^{\text{T}}$.

All simulation results are based on 1,000 Monte Carlo runs. Table 6.1 summarizes the simulation results with biases, standard errors, and coverage rates of 95% confidence intervals using asymptotic

normality of the point estimators. The following observations can be made from Table 6.1. $\hat{\mu}_{\text{ipw}}$ has large biases when the propensity score is misspecified. $\hat{\mu}_{\text{dr}}$ gains robustness over $\hat{\mu}_{\text{ipw}}$ if one of the outcome regression model or the propensity score is correctly specified. However, if both models are misspecified, $\hat{\mu}_{\text{dr}}$ has a larger bias. $\hat{\mu}_{\text{nni}}$ has small biases across four scenarios, which shows its robustness. Importantly, the performance of $\hat{\mu}_{\text{nni}}$ is close to that of $\hat{\mu}_{\text{HT}}$ in terms of standard errors and coverage rates, which is consistent with our theory in Theorem 1. Moreover, as predicted by our theoretical results, $\hat{\mu}_{\text{knn}}$ improves $\hat{\mu}_{\text{nni}}$ in terms of efficiency. Also, $\hat{\mu}_{\text{GAM}}$ shows robustness because of the flexibility of the model specification. The regression calibration estimator $\hat{\mu}_{\text{RC}}$ has small biases across all scenarios and therefore shows robustness against model specifications for sampling score and outcome. Moreover, it has smaller standard errors than both $\hat{\mu}_{\text{nni}}$ and $\hat{\mu}_{\text{knn}}$. The coverage rates are all close to the nominal level.

Table 6.1
Simulation results: bias, standard error, and coverage rate of 95% confidence intervals under four scenarios based on 1,000 Monte Carlo samples. OM: outcome model; PS: propensity score model (all numbers in the table are the numerical results multiplied by 100)

	Scenario I			Scenario II			Scenario III			Scenario IV		
OM	linear			linear			nonlinear			nonlinear		
PS	linear			nonlinear			linear			nonlinear		
	Bias	S.E.	C.R.	Bias	S.E.	C.R.	Bias	S.E.	C.R.	Bias	S.E.	C.R.
						Population Mean of Y_1						
$\hat{\mu}_{\text{HT}}$	0.2	6.5	96.0	-0.2	6.4	94.5	0.61	15.2	95.7	-0.5	15.6	93.5
$\hat{\mu}_{\text{ipw}}$	-0.1	1.6	95.3	22.2	35.8	97.5	-0.1	4.2	95.3	432.7	284.5	75.6
$\hat{\mu}_{\text{dr}}$	0.0	4.6	94.5	0.0	4.3	96.5	0.5	14.2	95.2	229.8	168.8	35.8
$\hat{\mu}_{\text{nni}}$	0.2	6.5	95.1	-0.3	6.4	94.7	0.7	15.2	94.6	-0.6	15.6	93.7
$\hat{\mu}_{\text{knn}}$	0.2	4.9	96.1	-0.3	4.9	95.6	0.5	14.5	94.6	-0.6	14.9	93.8
$\hat{\mu}_{\text{GAM}}$	0.1	4.5	95.7	-0.2	4.5	96.0	0.5	14.3	94.9	-0.6	14.8	93.4
$\hat{\mu}_{\text{RC}}$	0.0	3.2	95.5	-0.2	4.1	95.3	-0.1	4.8	95.0	0.1	6.7	95.5
						Population Mean of Y_2						
$\hat{\mu}_{\text{HT}}$	-0.0	1.5	96.2	-0.0	1.6	95.1	-0.1	1.6	95.2	0.1	1.6	94.4
$\hat{\mu}_{\text{ipw}}$	0.0	0.2	95.0	-12.1	3.1	0.0	-0.0	0.3	95.4	3.0	1.8	94.7
$\hat{\mu}_{\text{dr}}$	-0.0	0.9	95.0	-1.1	1.8	68.6	0.0	0.4	94.9	-2.9	2.2	59.8
$\hat{\mu}_{\text{nni}}$	0.0	1.4	95.3	-0.0	1.6	95.3	-0.1	1.6	94.6	0.1	1.6	95.3
$\hat{\mu}_{\text{knn}}$	0.0	1.0	95.8	-0.0	1.1	95.8	-0.0	1.0	95.2	0.0	0.9	96.1
$\hat{\mu}_{\text{GAM}}$	-0.0	0.9	95.3	-0.0	0.9	94.8	-0.0	0.8	96.2	0.0	0.8	94.5
$\hat{\mu}_{\text{RC}}$	0.0	1.2	95.5	-0.1	1.4	94.2	-0.0	1.4	94.1	0.1	1.5	95.6
						Conditional Mean of Y_1 given $Y_2 = 1$						
$\hat{\mu}_{\text{HT}}$	0.0	7.3	95.1	-0.3	7.2	95.2	0.2	9.3	95.3	-0.1	9.8	94.1
$\hat{\mu}_{\text{ipw}}$	-0.1	1.6	95.2	-9.1	10.3	69.8	-0.1	4.3	95.0	534.2	329.8	65.3
$\hat{\mu}_{\text{dr}}$	0.1	4.7	95.6	2.5	4.6	93.2	9.8	18.0	93.1	452.0	465.4	65.6
$\hat{\mu}_{\text{nni}}$	-0.0	7.3	95.0	-0.3	7.3	95.3	0.1	9.2	95.4	-2.2	9.5	95.2
$\hat{\mu}_{\text{knn}}$	-0.1	4.7	96.8	-0.3	4.6	96.5	0.1	6.0	94.8	0.0	6.4	93.6
$\hat{\mu}_{\text{GAM}}$	0.0	4.8	94.2	-0.3	4.5	96.0	-0.1	6.5	95.5	-0.6	6.8	94.8
$\hat{\mu}_{\text{RC}}$	-0.0	3.9	94.8	-0.2	5.0	96.0	-0.2	5.4	95.1	-0.1	5.4	96.7

6.2 Monthly retail trade survey

To demonstrate the practical relevance, we consider the U.S. Census Bureau's 2014 Monthly Retail Trade Survey (Mulry, Oliver and Kaputa, 2014). The Monthly Retail Trade Survey is an economic indicator survey whose monthly estimates are inputs to the Gross Domestic Product estimates. This survey selects a sample of about 12,000 retail businesses each month with paid employees to collect data on sales and inventories. It employs an one-stage stratified sample with stratification based on major industry, further substratified by the estimated annual sales referred to as the size variable.

For simulation purpose, we use the simulated data from the 2014 Monthly Retail Trade Survey to suggest the data generating model and the true parameter values (https://ww2.amstat.org/meetings/ices/2016/contests.cfm). We generate a finite population of N = 812,765 retail businesses with 16 strata with a stratum identifier h, sales Y, inventories $\mathbf{X}$, and a size variable Z on the log scale. Table 6.2 reports some summary statistics. We generate the inventory data from $X_{hi} \sim N\left(\mu_{X,h}, \sigma^2_{X,h}\right)$ for $i = 1, \ldots, N_h$ and $h = 1, \ldots, 16$, and the sales data from a linear model

$$Y_{hi} = \beta_0 + X_{hi} + \varepsilon_{hi}, \tag{6.5}$$

and a nonlinear model

$$Y_{hi} = \beta_0 + 0.5X^2_{hi} + \varepsilon_{hi}, \tag{6.6}$$

where $\varepsilon_{hi} \sim \mathcal{N}(0, 0.25)$. In (6.5) and (6.6), we specify different values for β_0 so that the parameter of interest, $\mu = N^{-1}\sum_{h=1}^{16}\sum_{i=1}^{N_h} Y_{hi}$, matches with the true population mean 12.73.

Table 6.2
The stratum size, sample allocation, mean and standard error of the inventory data on the log scale extracted from the 2014 Monthly Retail Trade simulated dataset

Stratum h	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
N_h	366	20	2,015	4,646	7,402	700	12,837	17,080	29,808	2,400	41,343	57,518	83,465	95,244	115,028	342,893
n_h	37	5	34	57	74	7	103	115	116	12	184	196	218	200	220	336
$\mu_{X,h}$	16.8	16.7	16.6	16.4	16.1	15.6	16.0	15.7	15.6	15.5	15.4	15.1	14.8	14.5	13.9	11.5
$\sigma_{X,h}$	1.1	0.8	0.4	0.3	0.4	0.6	0.4	0.4	0.4	0.3	0.4	0.4	0.3	0.7	0.5	1.1
$\mu_{Z,h}$	5.9	2.3	5.8	6.3	6.6	4.2	6.9	7.0	7.4	4.8	7.5	7.6	7.7	7.6	7.7	8.1

We also generate a big data sample $\mathcal{S}_B$ where the inclusion indicator $\delta_{hi} \sim \text{Ber}(p_{hi})$ with the inclusion probability p_{hi} for unit i in stratum h. The big data sample in practice is often available from E-commercial companies who monitor inventories and sales for retail businesses. For the big data inclusion probability, let $Z_{hi} \sim N\left(\mu_{Z,h}, \sigma^2_{Z,h}\right)$, for $i = 1, \ldots, N_h$ and $h = 1, \ldots, 16$. We consider a logistic linear model

$$\text{logit}(p_{hi}) = \alpha_0 + Z_{hi}, \tag{6.7}$$

and a nonlinear logistic model

$$\text{logit}(p_{hi}) = \alpha_0 + X_{hi} + Z_{hi}^2, \tag{6.8}$$

where we specify different values for α_0 so that the mean inclusion probability is about 30%. Lastly, we generate a representative sample $\mathcal{S}_A$ by stratified sampling with simple random sampling within strata without replacement; see Table 6.2 for the sample allocation.

We consider the seven estimators in Section 6.1 adopted for stratified sampling. In each mass imputed dataset, we apply the following point estimator and variance estimator: $\hat{\mu} = N^{-1}\sum_{h=1}^{H} N_h \bar{y}_{n_h}$ with $\bar{y}_{n_h}$ is the sample mean of y in the h^{th} stratum, $\hat{V}(\hat{\mu}) = N^{-2}\sum_{h=1}^{H} N_h^2 (1 - n_h / N_h) s_{n_h}^2 / n_h$ with $s_{n_h}^2 = (n_h - 1)^{-1}\sum_{i=1}^{n_h}(y_{hi} - \bar{y}_{n_h})^2$.

Table 6.3 summarizes the simulation results. A similar discussion to Section 6.1 applies. $\hat{\mu}_{\text{ipw}}$ is sensitive to misspecification of the selection model; while $\hat{\mu}_{\text{dr}}$ has double robustness feature, which still relies on at least one model to be correctly specified. Mass imputation based on nearest neighbor imputation, k nearest neighbor imputation and generalized additive model shows good performances by leveraging the representativeness of the survey sample and the predictive power of the big data sample. In addition, if the big data membership is known throughout the survey data, the regression calibration estimator gains efficiency while maintaining the robustness against model misspecification.

Table 6.3
Simulation results: bias, standard error, and coverage rate of 95% confidence intervals under four scenarios based on 1,000 Monte Carlo runs for the 2014 Monthly Retail Trade Survey. OM: outcome model; PS: propensity score model (all numbers in the table are the numerical results multiplied by 100)

OM	Scenario I linear			Scenario II linear			Scenario III nonlinear			Scenario IV nonlinear		
PS	linear			nonlinear			linear			nonlinear		
	Bias	S.E.	C.R.	Bias	S.E.	C.R.	Bias	S.E.	C.R.	Bias	S.E.	C.R.
$\hat{\mu}_{\text{HT}}$	0.0	3.0	95.0	0.0	3.0	95.0	1.1	31.5	95.0	1.1	31.5	95.0
$\hat{\mu}_{\text{ipw}}$	-0.6	5.8	96.6	-55.5	1.7	0.0	-7.3	76.2	96.6	-735.8	22.3	0.0
$\hat{\mu}_{\text{dr}}$	-0.3	2.7	94.4	-0.2	2.7	94.0	-3.3	34.6	93.8	-52.3	33.2	65.0
$\hat{\mu}_{\text{nni}}$	0.1	3.1	94.5	-0.1	3.1	94.6	1.1	31.5	95.3	-0.3	31.7	94.6
$\hat{\mu}_{\text{knn}}$	0.1	2.7	94.4	-0.2	2.7	94.3	1.0	31.4	94.9	-2.3	31.4	94.1
$\hat{\mu}_{\text{GAM}}$	0.1	2.7	94.9	0.1	2.7	94.9	1.1	31.6	94.9	-2.5	31.4	94.2
$\hat{\mu}_{\text{RC}}$	0.1	2.9	94.1	-0.1	2.6	95.1	0.6	30.7	94.6	-0.5	26.9	95.0

7. Real-data application

7.1 Data description

To demonstrate the practical use, we apply the proposed method to the survey data from the Korea National Health and Nutrition Examination Survey (KNHANES) and the big data from National Health Insurance Sharing Service (NHISS). The KNHANES is an annual national survey that studies the health and nutritional status of Koreans since 1998. The surveys have been conducted by the Korea Centers for

Disease Control and Prevention. This nationally representative cross-sectional survey includes approximately 10,000 individuals each year as a survey sample and collects information on socioeconomic status, health-related behaviours, quality of life, healthcare utilization, anthropometric measures, biochemical and clinical profiles for non-communicable diseases and dietary intakes with three component surveys: health interview, health examination, and nutrition survey. More details of the KNHANES can be found in Kweon, Kim, jin Jang, Kim, Kim, Choi, Chun, Khang and Oh (2014). The data set used in this study has 4,929 samples.

On the other hand, the big data from NHISS provides health-related information collected from National Health Screening Program (NHSP) in South Korea. The NHSP was launched with the goal of improving the overall health of the South Korean citizens and preventing the costly chronic diseases. All beneficiaries are eligible for screening once every year or two depending on their demographic or occupational status. The specific screening items are stipulated by the implementation standards, which include, but not limited to, various blood tests and cancer screening. The total number of eligible beneficiaries is about 16 million, where approximately 75% of them participated the screening. The data that we have used in this study is the subset corresponding to the blood test results that are associated with metabolic syndrome from the 2014 program. The variables in this data set are demographics as sex and age, and clinical measurements such as total glycerides (mg/dL), total cholesterol (mg/dL), high-density lipoprotein cholesterol (HDL, mg/dL), and medical diagnosis on whether having anemia. The data set is made publicly available after anonymization and randomly selecting 1 million observations (National Health Insurance Data Sharing Service, 2014). Note that more thorough data can be purchased with a paid subscription and expert panel review.

7.2 Analysis and results

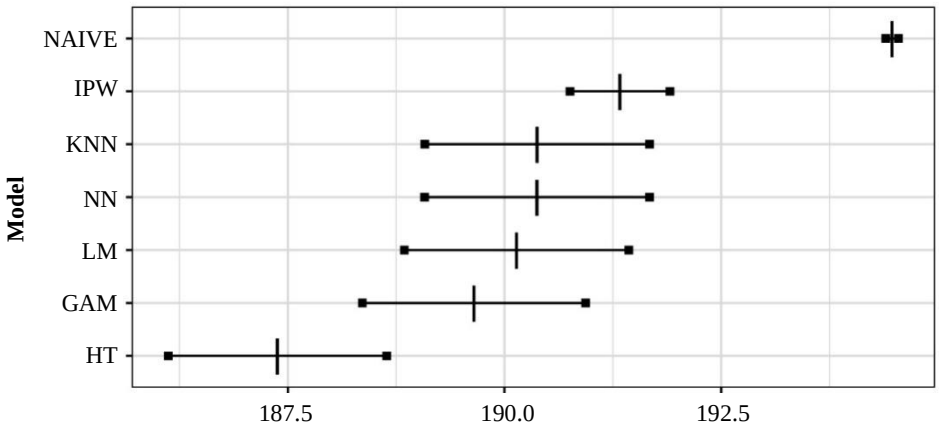
To apply the proposed method of mass imputation, we assume that total cholesterol is not available in KNHANES data, and use the big data from NHSP to perform mass imputation for total cholesterol variable. The actual survey values from KNHANES are used to compute a benchmark so that we can validate the efficacy of our proposed method. We consider the six different estimators:

- HT: the Horvitz-Thompson estimator based on the Sample A data. This is used for a benchmark comparison;
- NN: the nearest neighbor imputation estimator;
- kNN: the k nearest neighbor imputation estimator with $k = 5$;
- GAM: the generalized additive model imputation estimator;
- LM: the linear regression model imputation estimator using sex, age group, HDL cholesterol and total glycerides as the covariates;
- IPW: the inverse propensity score weighting estimator;
- NAIVE: the naive estimator using the Sample B without any treatment.

Total cholesterol is affected by the amount of HDL, because HDL is one of the components that constitute the total cholesterol, and is known to be also affected by sex and age. Unless Sample B is from

a particular sub-population such as cardiovascular stenosis patients group, we may assume that the relationship between the total cholesterol and other variables remain the same. Hence ignorability holds. Also the covariates are all medical/biological measurements, meaning they should stay within the similar range both for Samples A and B. The variance estimator for each estimator is calculated, and 95% Wald confidence interval for μ_g is obtained using asymptotic normality. Figure 7.1 depicts the intervals, where the population mean estimate from each method is presented as a vertical bar. The interval obtained from HT can be viewed as a reference. It can be seen that all estimators produce intervals that are slightly overestimated compared to the one from HT. It is because of the inherent bias in total cholesterol level in NHSP data; the sample mean values of the total cholesterol from NHSP data is 7 point or 3.7 per cent higher than HT estimator calculated with KNHANES data, as seen from the naive result. One can see that all the proposed methods substantially reduce such bias to make the estimator close to the HT estimator, which shows the benefit of the proposed methods. IPW estimator produced a relatively poor estimate compared to other methods, which is probably because of either misspecification in logistic regression model, or considerable discrepancy in sample sizes between KNHANES (4,929) and NHSP data (1 million). We also tried the DR estimator but not included here, because the effect from the IPW is very marginal due to the limited auxiliary variables available.

Figure 7.1 Estimated 95% confidence interval for the total cholesterol level.



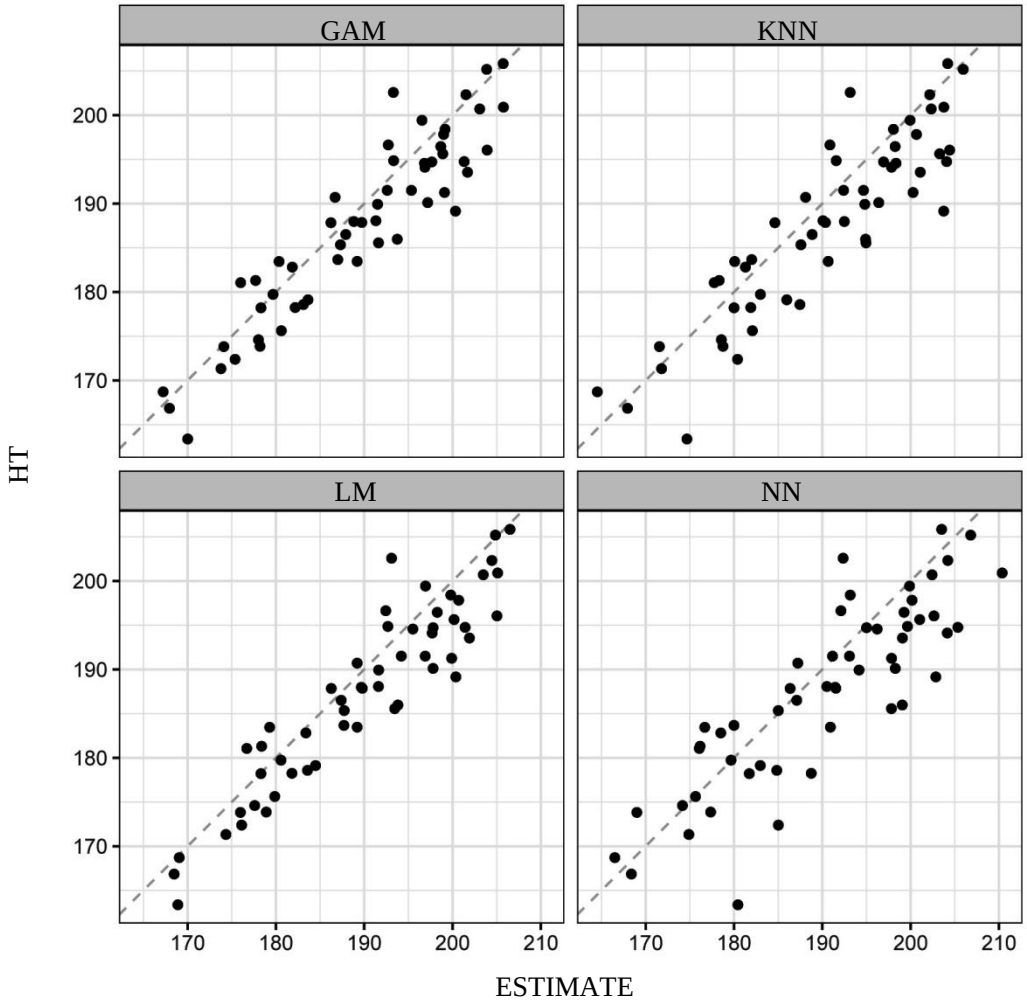
To better understand the prediction performance of the mass imputation methods, we calculated RMSE, mean bias, and correlation of imputed values by comparing the imputed values and actual survey values. Because we can observe the actual survey values from KNHANES, we can compute the prediction quality measures. Table 7.1 presents the summarized table, where we compared the results at individual levels and subgroup mean levels divided by age group and sex. For subgroup levels, we first obtain the subgroup mean estimates and then calculate the statistics aggregated over different groups. It can be seen that GAM performs better than the other methods in terms of RMSE and correlation. Overall, mass imputation method provides reasonable results for subgroup level as can be seen in Figure 7.2.

These quality measures need a predicted value for Sample A, hence IPW estimators are excluded in the comparison. Estimating the population and subgroup means using Sample B can give a very biased result – in the case of NHSP data, the difference between the mean of NHSP data and the HT estimator from KNHANES is about 7.09, or 3.7 per cent.

Table 7.1
Comparison of the imputation methods

	Method	RMSE	Bias	Corr.
Individual	NN	43.94	2.87	0.26
	KNN	32.62	2.86	0.42
	GAM	29.15	2.13	0.54
	LM	30.35	2.59	0.48
Group Means	NN	6.33	2.68	0.85
	KNN	5.44	2.70	0.90
	GAM	4.33	2.03	0.93
	LM	4.57	2.52	0.93

Figure 7.2 Comparison of the HT estimates and estimates using mass imputation for subgroup average.



8. Discussion

Mass imputation is an important technique for survey data integration. When the training dataset for imputation is obtained from a probability sample, the theory of Kim and Rao (2012) can be directly applied. If the training dataset is a non-probability sample and its size is huge, we have shown in this paper that various non-parametric methods can be used for mass imputation, and the estimation error in the imputation model can be safely ignored, under the assumption that the sampling mechanism for training data is missing at random in the sense of Rubin (1976). If the sampling mechanism is believed to be missing not at random, imputation techniques can be developed under the strong structural assumptions for the sampling mechanism (e.g., Riddles, Kim and Im, 2016; Morikawa and Kim, 2018) or the outcome model (e.g., Yang, Zeng and Wang, 2020). Also, when the training dataset has a hierarchical structure, multi-level models can be used to develop mass imputation. This is closely related to unit-level small area estimation in survey sampling (Rao and Molina, 2015).

The mass imputation estimator is not necessarily efficient. In Section 5, we have described a method of using calibration weighting as a tool for efficient data integration with big data. The calibration weighting requires correct matching between two data sources, as investigated by Kim and Tam (2020). Also, if the fraction of big data in the finite population is not substantial, the efficiency gain will be limited. Instead, one could improve the efficiency by combining the mass imputation estimator with the inverse propensity weighting estimator in the big data (Yang, Kim and Song, 2020). However, the correct specification of the propensity score model will be challenging. These are topics for future research.

Acknowledgements

We thank two anonymous referees and the associated editor for very constructive comments. Dr. Yang is partially supported by NSF grant DMS 1811245 and NIA grant 1R01AG066883. Dr. Kim is partially supported by NSF grant MMS 1733572 and the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa.

Appendix

A.1 Proof for Theorem 1

For a given $\mathbf{X}_i = \mathbf{x}$ in Sample A, we show that $\mathbf{X}_{i(1)}$ converges to $\mathbf{x}$ in probability as $N_B \to \infty$. Consider for any $\varepsilon > 0$, we show that

$$P\left\{d\left(\mathbf{X}_{i(1)}, \mathbf{x}\right) > \varepsilon\right\} = P\left\{d\left(\mathbf{X}_j, \mathbf{x}\right) > \varepsilon, \forall j \in B\right\} \tag{A.1}$$

converges to zero, and therefore $\mathbf{X}_{i(1)}$ converges to $\mathbf{x}$ in probability as $N_B \to \infty$, where the probability is induced by the sampling process of Sample B of size N_B. We show this fact by contradiction. Assume that for some $\varepsilon > 0$, $P\left\{d\left(\mathbf{X}_{i(1)}, \mathbf{x}\right) > \varepsilon\right\}$ does not coverage to zero as $N_B \to \infty$. Define the region $\mathcal{R}_\varepsilon = \{\mathbf{X}: d(\mathbf{X}, \mathbf{x}) \le \varepsilon\}$. Then, we must have $f(\mathbf{X} \mid \delta_B = 1) = 0$ for $\mathbf{X} \in \mathcal{R}_\varepsilon$; otherwise, there exists $\tilde{\mathbf{X}} \in \mathcal{R}_\varepsilon$ with a positive probability in Sample B as $N_B \to \infty$, and therefore $P\left\{d\left(\mathbf{X}_{i(1)}, \mathbf{x}\right) > \varepsilon\right\} = 0$ as $N_B \to \infty$. But the claim that $f(\mathbf{X} \mid \delta_B = 1) = 0$ for $\mathbf{X} \in \mathcal{R}_\varepsilon$ implies that $\mathcal{R}_\varepsilon$ is a non-overlap region of the distribution of $\mathbf{X}$ between Sample A (and also the population) and Sample B, violating Assumption 2.

Given $\mathbf{X}_i = \mathbf{x}$ in Sample A, for any continuous and bounded $g(y)$,

$$\begin{aligned} E\{g(Y_{i(1)}) | \mathbf{X}_i = \mathbf{x}, i \in A\} &= E\left[E\{g(Y_{i(1)}) | \mathbf{X}_{i(1)}, \mathbf{X}_i = \mathbf{x}, i \in A\} | \mathbf{X}_i = \mathbf{x}, i \in A\right] \\ &= E\left[E\{g(Y_{i(1)}) | \mathbf{X}_{i(1)}\} | \mathbf{X}_i = \mathbf{x}, i \in A\right] \\ &= E\{\mu_g(\mathbf{X}_{i(1)}) | \mathbf{X}_i = \mathbf{x}, i \in A\} \to E\{\mu_g(\mathbf{X}_i) | \mathbf{X}_i = \mathbf{x}, i \in A\} \\ &= E\{g(Y_i) | \mathbf{X}_i = \mathbf{x}, i \in A\}, \end{aligned}$$

in probability as $N_B \to \infty$, where $\to$ follows from the fact that $\mu_g(\mathbf{x})$ is bounded and continuous. Then, by Portmanteau Lemma (Klenke, 2006), $Y_{i(1)} \to Y_i | (\mathbf{X}_i = \mathbf{x}, i \in A)$ in distribution as $N_B \to \infty$. By Assumption 1, $g(Y_{i(1)}) | (\mathbf{X}_i, i \in A) \to \mu_g(\mathbf{X}_i) + e_g^*(\mathbf{X}_i)$ in distribution as $N_B \to \infty$, where $e_g^*(\mathbf{X}_i)$ has the same distribution as $\{g(Y_i) | (\mathbf{X}_i, i \in A)\} - \mu_g(\mathbf{X}_i)$.

We now show that for $i \neq j \in A$, $e_g^*(\mathbf{X}_i)$ and $e_g^*(\mathbf{X}_j)$ are conditionally independent, given data $\mathcal{O}_A$. It is sufficient to show that $P\{i(1) = j(1)\} \to 0$ as $N_B \to \infty$; in other words, the same unit can not be matched for unit i and unit j with probability 1. This can be shown using (A.1) with $\varepsilon = \min_{i \neq j \in A} \| \mathbf{X}_i - \mathbf{X}_j \|$.

Therefore, conditional on data $\mathcal{O}_A$, we have

$$\hat{\mu}_{g,\text{nni}} = \frac{1}{N} \sum_{i \in A} \pi_i^{-1} g(Y_{i(1)}) \to \frac{1}{N} \sum_{i \in A} \pi_i^{-1} g(Y_i) = \hat{\mu}_{g,\text{HT}}$$

in distribution as $N_B \to \infty$. This completes the proof for Theorem 1.

Let

$$\tilde{V}_{\text{nni}} = \frac{n}{N^2} \sum_{i \in A} \sum_{j \in A} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \frac{g(Y_i)}{\pi_i} \frac{g(Y_j)}{\pi_j}. \tag{A.2}$$

Then, $\tilde{V}_{\text{nni}}$ is consistent for V_{nni}.

Similar to the above argument, for $i, j \in A$, conditional on data $\mathcal{O}_A$, $g(Y_{i(1)}) g(Y_{j(1)}) \to g(Y_i) g(Y_j)$ as $N_B \to \infty$. Therefore, conditional on data $\mathcal{O}_A$,

$$\hat{V}_{\text{nni}} = \frac{n}{N^2} \sum_{i \in A} \sum_{j \in A} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \frac{g(Y_{i(1)})}{\pi_i} \frac{g(Y_{j(1)})}{\pi_j} \to \tilde{V}_{\text{nni}}, \tag{A.3}$$

in distribution as $N_B \to \infty$. Combining (A.2) and (A.3), $\hat{V}_{\text{nni}}$ is consistent for V_{nni}.

A.2 Proof for Theorem 2

To investigate the asymptotic properties of $\hat{\mu}_{g,\text{knn}}$, we re-express

$$\hat{\mu}_g(\mathbf{x}) = \frac{\sum_{j \in B} K_{R_\mathbf{x}}(\mathbf{x} - \mathbf{X}_j) g(Y_j)}{\sum_{j \in B} K_{R_\mathbf{x}}(\mathbf{x} - \mathbf{X}_j)},$$

where

$$K_h(u) = \frac{1}{h^p} K\left(\frac{u}{h}\right), \quad K(u) = 0.5I(\|u\| \le 1),$$

and the bandwidth $h = R_{\mathbf{x}}$ is the random distance between $\mathbf{x}$ and its furthest among the k nearest neighbors. Therefore, $\hat{\mu}_{g,\text{knn}}$ can be viewed as a kernel estimator incorporating a data-driven bandwidth.

In the literature, asymptotic properties of the k nearest neighbor imputation estimator have been studied extensively. The result shown in the following lemma on k nearest neighbor imputation is extracted from Mack (1981).

Lemma 1. *Under Assumptions 1-3,*

$$N^{-1}\sum_{j=1}^{N} \delta_{B,j} K_{R_{\mathbf{x}}}(\mathbf{x} - \mathbf{X}_j) g(Y_j) = f(\mathbf{x})\pi_B(\mathbf{x})\mu_g(\mathbf{x}) + O_{\text{p}}\left\{\left(\frac{k}{N}\right)^{2/p} + \frac{1}{k}\right\}. \tag{A.4}$$

We now express

$$\hat{\mu}_{g,\text{knn}} = \frac{1}{N}\sum_{i=1}^{N} \pi_i^{-1}\delta_{A,i}\mu_g(\mathbf{X}_i) + \frac{1}{N}\sum_{i=1}^{N} \pi_i^{-1}\delta_{A,i}\{\hat{\mu}_g(\mathbf{X}_i) - \mu_g(\mathbf{X}_i)\}.$$

Let $T_N = N^{-1}\sum_{i=1}^{N} \pi_i^{-1}\delta_{A,i}\{\hat{\mu}_g(\mathbf{X}_i) - \mu_g(\mathbf{X}_i)\}$. To study the properties for T_N, we first look at $\hat{\mu}_g(\mathbf{x})$, which can be expressed as

$$\hat{\mu}_g(\mathbf{x}) = \frac{h_N(\mathbf{x})}{f_N(\mathbf{x})},$$

where $h_N(\mathbf{x}) \equiv N^{-1}\sum_{j=1}^{N} \delta_{B,j} K_{R_{\mathbf{x}}}(\mathbf{x} - \mathbf{X}_j) g(Y_j)$ and $f_N(\mathbf{x}) \equiv N^{-1}\sum_{j=1}^{N} \delta_{B,j} K_{R_{\mathbf{x}}}(\mathbf{x} - \mathbf{X}_j)$. By the result in Lemma 1, we obtain

$$h_N(\mathbf{x}) = f(\mathbf{x})\pi_B(\mathbf{x})\mu_g(\mathbf{x}) + O_{\text{p}}\left\{\left(\frac{k}{N}\right)^{2/p} + \frac{1}{k}\right\}$$

$$f_N(\mathbf{x}) = f(\mathbf{x})\pi_B(\mathbf{x}) + O_{\text{p}}\left\{\left(\frac{k}{N}\right)^{2/p} + \frac{1}{k}\right\}.$$

Now, by a Taylor expansion, we obtain

$$\begin{aligned}
\hat{\mu}_g(\mathbf{x}) - \mu_g(\mathbf{x}) &= \frac{h_N(\mathbf{x})}{f_N(\mathbf{x})} - \mu_g(\mathbf{x}) \\
&= \frac{1}{f(\mathbf{x})\pi_B(\mathbf{x})}\{h_N(\mathbf{x}) - f(\mathbf{x})\pi_B(\mathbf{x})\mu_g(\mathbf{x})\} \\
&\quad - \frac{f(\mathbf{x})\pi_B(\mathbf{x})\mu_g(\mathbf{x})}{\{f(\mathbf{x})\pi_B(\mathbf{x})\}^2}\{f_N(\mathbf{x}) - f(\mathbf{x})\pi_B(\mathbf{x})\} + O_{\text{p}}\left\{\left(\frac{k}{N}\right)^{2/p} + \frac{1}{k}\right\} \\
&= \frac{1}{f(\mathbf{x})\pi_B(\mathbf{x})}\{h_N(\mathbf{x}) - f_N(\mathbf{x})\mu_g(\mathbf{x})\} + O_{\text{p}}\left\{\left(\frac{k}{N}\right)^{2/p} + \frac{1}{k}\right\}.
\end{aligned}$$

Therefore, we obtain

$$T_N = \frac{1}{N^2}\sum_{i=1}^{N}\frac{\delta_{A,i}}{\pi_i}\frac{1}{f(\mathbf{X}_i)\pi_B(\mathbf{X}_i)}\sum_{j=1}^{N}\delta_{B,j}K_{R_{\mathbf{X}_i}}(\mathbf{X}_i - \mathbf{X}_j)\{g(Y_j) - \mu_g(\mathbf{X}_i)\} + O_{\mathrm{p}}\left\{\left(\frac{k}{N}\right)^{2/p} + \frac{1}{k}\right\}.$$

Under the assumption in Theorem 2, it is easy to derive that $(k/N)^{2/p} + 1/k = o(n^{-1/2})$, and therefore,

$$T_N = \frac{1}{N^2}\sum_{i=1}^{N}\frac{\delta_{A,i}}{\pi_i}\frac{1}{f(\mathbf{X}_i)\pi_B(\mathbf{X}_i)}\sum_{j=1}^{N}\delta_{B,j}K_{R_{\mathbf{X}_i}}(\mathbf{X}_i - \mathbf{X}_j)\{g(Y_j) - \mu_g(\mathbf{X}_i)\} + o_{\mathrm{p}}(n^{-1/2}).$$

We then express T_N in a form of U-statistics (van der Vaart, 2000; Chapter 12):

$$T_N = \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j\neq i}h(\mathbf{Z}_i, \mathbf{Z}_j) + o_{\mathrm{p}}(n^{-1/2}),$$

where $\mathbf{Z}_i = (\mathbf{X}_i, Y_i, \delta_{A,i}, \delta_{B,i})$ and

$$\begin{aligned} h(\mathbf{Z}_i, \mathbf{Z}_j) &= \frac{1}{2}\left[\frac{\delta_{A,i}\delta_{B,j}}{\pi_i}\frac{1}{f(\mathbf{X}_i)\pi_B(\mathbf{X}_i)}K_{R_{\mathbf{X}_i}}(\mathbf{X}_i - \mathbf{X}_j)\{g(Y_j) - \mu_g(\mathbf{X}_i)\}\right. \\ &\quad \left. + \frac{\delta_{A,j}\delta_{B,i}}{\pi_j}\frac{1}{f(\mathbf{X}_j)\pi_B(\mathbf{X}_j)}K_{R_{\mathbf{X}_j}}(\mathbf{X}_j - \mathbf{X}_i)\{g(Y_i) - \mu_g(\mathbf{X}_j)\}\right] \\ &\equiv \frac{1}{2}(\zeta_{ij} + \zeta_{ji}). \end{aligned}$$

Now, by Lemma 1, we obtain

$$\begin{aligned} E(\zeta_{ij}\,|\,\mathbf{Z}_i) &= E\left[\frac{\delta_{A,i}\delta_{B,j}}{\pi_i}\frac{1}{f(\mathbf{X}_i)\pi_B(\mathbf{X}_i)}K_{R_{\mathbf{X}_i}}(\mathbf{X}_i - \mathbf{X}_j)\{g(Y_j) - \mu_g(\mathbf{X}_i)\}\,\middle|\,\mathbf{Z}_i\right] \\ &= \frac{\delta_{A,i}}{\pi_i}\frac{1}{f(\mathbf{X}_i)\pi_B(\mathbf{X}_i)}E\left[\delta_{B,j}K_{R_{\mathbf{X}_i}}(\mathbf{X}_i - \mathbf{X}_j)\{g(Y_j) - \mu_g(\mathbf{X}_i)\}\,\middle|\,\mathbf{Z}_i\right] \\ &= O_{\mathrm{p}}\left\{\left(\frac{k}{N}\right)^{2/p} + \frac{1}{k}\right\}, \end{aligned}$$

and

$$\begin{aligned} E(\zeta_{ji}\,|\,\mathbf{Z}_i) &= E\left[\frac{\delta_{A,j}\delta_{B,i}}{\pi_j}\frac{1}{f(\mathbf{X}_j)\pi_B(\mathbf{X}_j)}K_{R_{\mathbf{X}_j}}(\mathbf{X}_j - \mathbf{X}_i)\{g(Y_i) - \mu_g(\mathbf{X}_j)\}\,\middle|\,\mathbf{Z}_i\right] \\ &= \delta_{B,i}E\left(E\left[\frac{\delta_{A,j}}{\pi_j}\frac{1}{f(\mathbf{X}_j)\pi_B(\mathbf{X}_j)}K_{R_{\mathbf{X}_j}}(\mathbf{X}_j - \mathbf{X}_i)\{g(Y_i) - \mu_g(\mathbf{X}_j)\}\,\middle|\,R_{\mathbf{X}_j}, \mathbf{Z}_i\right]\,\middle|\,\mathbf{Z}_i\right) \\ &= \frac{\delta_{B,i}}{\pi_B(\mathbf{X}_i)}\{g(Y_i) - \mu_g(\mathbf{X}_i)\} + O_{\mathrm{p}}\left\{\left(\frac{k}{N}\right)^{2/p} + \frac{1}{k}\right\}. \end{aligned}$$

Therefore, by the theory of U-statistics, we obtain

$$\begin{aligned} T_N &= \frac{2}{N}\sum_{i=1}^{N} E\left\{h\left(\mathbf{Z}_i, \mathbf{Z}_j\right)\middle|\mathbf{Z}_i\right\} + o_{\mathrm{p}}\left(n^{-1/2}\right) \\ &= \frac{1}{N}\sum_{i=1}^{N}\frac{\delta_{B,i}}{\pi_B\left(\mathbf{X}_i\right)}\left\{g\left(Y_i\right) - \mu_g\left(\mathbf{X}_i\right)\right\} + o_{\mathrm{p}}\left(n^{-1/2}\right). \end{aligned}$$

Combining the above results leads to

$$\begin{aligned} \hat{\mu}_{g,\mathrm{knn}} - \mu_g &= \frac{1}{N}\sum_{i=1}^{N}\left\{\pi_i^{-1}\delta_{A,i}\mu_g\left(\mathbf{X}_i\right) - \mu_g\left(\mathbf{X}_i\right)\right\} \\ &\quad + \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{\delta_{B,i}}{\pi_B\left(\mathbf{X}_i\right)} - 1\right\}\left\{g\left(Y_i\right) - \mu_g\left(\mathbf{X}_i\right)\right\} + o_{\mathrm{p}}\left(n^{-1/2}\right). \end{aligned} \tag{A.5}$$

Then, the asymptotic results in Theorem 2 follow by Assumptions 1-4 and (A.5).

A.3 Proof for Theorem 3

The consistency and asymptotic normality of $n^{1/2}\hat{\mu}_{g,\mathrm{nni}}$ follow by the standard arguments under Assumptions 1-4. The remaining is to show that the asymptotic variance of $n^{1/2}\hat{\mu}_{g,\mathrm{nni}}$ is V_{nni}.

Using the distance function $G\left(\omega_i, d_i\right) = d_i\left(\omega_i/d_i - 1\right)^2$ in (5.1), the minimum distance estimation leads to generalized regression estimation (Park and Fuller, 2012). Therefore, we express

$$\begin{aligned} n^{1/2}\hat{\mu}_g &= \frac{n^{1/2}}{N}\sum_{i\in A}\omega_i g\left(Y_{i(1)}\right) \\ &= \frac{n^{1/2}}{N}\sum_{i\in A}\pi_i^{-1}g\left(Y_{i(1)}\right) - \frac{n^{1/2}}{N}\left(\sum_{i\in A}\pi_i^{-1}\mathbf{h}_i^{*\mathrm{T}}\boldsymbol{\beta}_N - \sum_{i=1}^{N}\mathbf{h}_i^{*\mathrm{T}}\boldsymbol{\beta}_N\right) + o_{\mathrm{p}}\left(n^{-1/2}\right). \end{aligned} \tag{A.6}$$

Similar to the argument in the proof for Theorem 1, we express

$$\begin{aligned} n^{1/2}\hat{\mu}_g &= \frac{n^{1/2}}{N}\sum_{i\in A}\pi_i^{-1}g\left(Y_{i(1)}\right) - \frac{n^{1/2}}{N}\left(\sum_{i\in A}\pi_i^{-1}\mathbf{h}_i^{*\mathrm{T}}\boldsymbol{\beta}_N - \sum_{i=1}^{N}\mathbf{h}_i^{*\mathrm{T}}\boldsymbol{\beta}_N\right) + o_{\mathrm{p}}\left(n^{-1/2}\right) \\ &= \frac{n^{1/2}}{N}\sum_{i\in A}\pi_i^{-1}\left\{g\left(Y_{i(1)}\right) - \mathbf{h}_i^{*\mathrm{T}}\boldsymbol{\beta}_N\right\} + \frac{n^{1/2}}{N}\sum_{i=1}^{N}\mathbf{h}_i^{*\mathrm{T}}\boldsymbol{\beta}_N + o_{\mathrm{p}}\left(n^{-1/2}\right). \end{aligned} \tag{A.7}$$

It is straightforward to show the variance of the second term in (A.7) is negligible given $nN^{-1} = o(1)$. Following the arguments in the proof for Theorems 1 and 2, $g\left(Y_{i(1)}\right)$ and $\mathbf{h}_i^*$ have the asymptotic distribution as $g\left(Y_i\right)$ and $\mathbf{h}_i$ given the data $\mathcal{O}_A$ from Sample A, respectively. Therefore, the asymptotic variance of $n^{1/2}\hat{\mu}_g$ is

$$V_{\mathrm{RC}} = \lim_{n\to\infty}\mathrm{var}\left[\frac{n^{1/2}}{N}\sum_{i\in A}\pi_i^{-1}\left\{g\left(Y_i\right) - \mathbf{h}_i^{\mathrm{T}}\boldsymbol{\beta}_N\right\}\right].$$

References

Abadie, A., and Imbens, G.W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74, 235-267.

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

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/pub/12-001-x/2020001/article/00001-eng.pdf.

Belloni, A., Chernozhukov, V., Chetverikov, D. and Kato, K. (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186, 345-366.

Beretta, L., and Santaniello, A. (2016). Nearest neighbor imputation algorithms: A critical evaluation. *BMC Medical Informatics and Decision Making*, 16, 198-208.

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

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

Breidt, F.J., McVey, A. and Fuller, W.A. (1996). Two-phase estimation by imputation. *Journal of the Indian Society of Agricultural Statistics*, 49, 79-90.

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

Brookhart, M.A., Schneeweiss, S., Rothman, K.J., Glynn, R.J., Avorn, J. and Stürmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, 163, 1149-1156.

Buchanan, A.L., Hudgens, M.G., Cole, S.R., Mollan, K.R., Sax, P.E., Daar, E.S., Adimora, A.A., Eron, J.J. and Mugavero, M.J. (2018). Generalizing evidence from randomized trials using inverse probability of sampling weights. *Journal of the Royal Statistical Society, Series A, (Statistics in Society)*, 181, 1193-1209.

Buelens, B., Burger, J. and van den Brakel, J.A. (2018). Comparing inference methods for nonprobability samples. *International Statistical Review*, 86, 322-343.

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

Cheng, P.E. (1994). Nonparametric estimation of mean functionals with data missing at random. *Journal of American Statistical Association*, 89, 81-87.

Chipperfield, J., Chessman, J. and Lim, R. (2012). Combining household surveys using mass imputation to estimate population totals. *Australian & New Zealand Journal of Statistics*, 54, 223-238.

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

Cochran, W.G. (2007). *Sampling Techniques*, New York: John Wiley & Sons, Inc.

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

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

DiSogra, C., Cobb, C., Chan, E. and Dennis, J.M. (2011). Calibrating non-probability internet samples with probability samples using early adopter characteristics. *Joint Statistical Meetings (JSM), Survey Research Methods*, 4501-4515.

Dong, L., Yang, S., Wang, X., Zeng, D. and Cai, J. (2020). Integrative analysis of randomized clinical trials with real world evidence studies, *arXiv preprint arXiv:2003.01242*.

Eilers, P.H., and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11, 89-102.

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

Fuller, W.A. (2009). *Sampling Statistics*, Wiley, Hoboken.

Hastie, T., and Tibshirani, R. (1990). *Generalized Additive Models*, NY: Chapman and Hall, Inc.

Imbens, G., Johnson, P. and Spady, R.H. (1998). Information theoretic approaches to inference in moment condition models. *Econometrica*, 66, 333-357.

Kalton, G., and Kish, L. (1984). Some efficient random imputation methods. *Communications in Statistics-Theory and Methods*, 13, 1919-1939.

Kang, J.D., and Schafer, J.L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22, 523-539.

Keiding, N., and Louis, T.A. (2016). Perils and potentials of self-selected entry to epidemiological studies and surveys. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 179, 319-376.

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

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

Kim, J.K., and Tam, S.M. (2020). Data integration by combining big data and survey sample data for finite population inference. *International Statistical Review*. Accepted (Available at https://doi.org/10.1111/insr.12434).

Kim, J.K., and Wang, Z. (2019). Sampling techniques for big data analysis. *International Statistical Review*, 87, 177-191.

Kitamura, Y., and Stutzer, M. (1997). An information-theoretic alternative to generalized method of moments estimation. *Econometrica*, 65, 861-874.

Klenke, A. (2006). *Probability Theory*, Springer-Verlag: Heidelberg.

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

Kweon, S., Kim, Y., jin Jang, M., Kim, Y., Kim, K., Choi, S., Chun, C., Khang, Y.-H. and Oh, K. (2014). Data resource profile: The Korea National Health and Nutrition Examination Survey (KNHANES). *International Journal of Epidemiology*, 43, 69-77.

Lee, S., and Valliant, R. (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods & Research*, 37, 319-343.

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

Mack, Y.-P. (1981). Local properties of k-NN regression estimates. *SIAM Journal on Algebraic Discrete Methods*, 2, 311-323.

Mack, Y., and Rosenblatt, M. (1979). Multivariate k-nearest neighbor density estimates. *Journal of Multivariate Analysis*, 9, 1-15.

McCullagh, P. (1984). Generalized linear models. *European Journal of Operational Research*, 16, 285-292.

McRoberts, R.E., Tomppo, E.O. and Næsset, E. (2010). Advances and emerging issues in national forest inventories. *Scandinavian Journal of Forest Research*, 25, 368-381.

Morikawa, K., and Kim, J.K. (2018). A note on the equivalence of two semiparametric estimation methods for nonignorable nonresponse. *Statistics & Probability Letters*, 140, 1-6.

Mulry, M.H., Oliver, B.E. and Kaputa, S.J. (2014). Detecting and treating verified influential values in a Monthly Retail Trade Survey. *Journal of Official Statistics*, 30, 721-747.

National Health Insurance Data Sharing Service (2014). National health screening data. https://nhiss.nhis.or.kr/bd/ab/bdabf006cv.do, [Accessed: 2019-07-11].

Nelder, J.A., and Baker, R.J. (1972). *Generalized Linear Models*, New York: John Wiley & Sons, Inc.

Newey, W.K., and Smith, R.J. (2004). Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica*, 72, 219-255.

O'Muircheartaigh, C., and Hedges, L.V. (2014). Generalizing from unrepresentative experiments: A stratified propensity score approach. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 63, 195-210.

Palmer, J.R., Espenshade, T.J., Bartumeus, F., Chung, C.Y., Ozgencil, N.E. and Li, K. (2013). New approaches to human mobility: Using mobile phones for demographic research. *Demography*, 50, 1105-1128.

Park, M., and Fuller, W.A. (2012). Generalized regression estimators. *Encyclopedia of Environmetrics*, 2, 1162-1166.

Pfeffermann, D., Eltinge, J.L. and Brown, L.D. (2015). Methodological issues and challenges in the production of official statistics: 24th Annual Morris Hansen Lecture. *Journal of Survey Statistics and Methodology*, 3, 425-483.

Rao, J.N.K. (2020). On making valid inferences by integrating data from surveys and other sources. *Sankhyā B*. pp. DOI 10.1007/s13571–020–00227–w.

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

Riddles, M.K., Kim, J.K. and Im, J. (2016). A propensity-score-adjustment method for nonignorable nonresponse. *Journal of Survey Statistics and Methodology*, 4, 215-245.

Rivers, D. (2007). Sampling for web surveys. *Joint Statistical Meetings*.

Rosenbaum, P.R., and Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

Ruppert, D., Wand, M.P. and Carroll, R.J. (2009). Semiparametric regression during 2003-2007. *Electronic Journal of Statistics*, 3, 1193-1256.

Särndal, C.-E., Swensson, B. and Wretman, J. (2003). *Model-Assisted Survey Sampling*. Springer Science & Business Media, New York: Springer-Verlag.

Schennach, S.M. (2007). Point estimation with exponentially tilted empirical likelihood. *Annals of Statistics*, 35, 634-672.

Stuart, E.A., Bradshaw, C.P. and Leaf, P.J. (2015). Assessing the generalizability of randomized trial results to target populations. *Prevention Science*, 16, 475-485.

Stuart, E.A., Cole, S.R., Bradshaw, C.P. and Leaf, P.J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 174, 369-386.

Tam, S.-M., and Clarke, F. (2015). Big data, official statistics and some initiatives by the Australian Bureau of Statistics. *International Statistical Review*, 83, 436-448.

Tam, S.-M., and Kim, J.-K. (2018). Big data ethics and selection-bias: An official statistician's perspective. *Statistical Journal of the IAOS*, 34, 577-588.

Tourangeau, R., Conrad, F.G. and Couper, M.P. (2013). *The Science of Web Surveys*, New York: Oxford University Press.

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

van der Vaart, A.W. (2000). *Asymptotic Statistics*, Cambridge University Press, Cambridge, MA.

Vavreck, L., and Rivers, D. (2008). The 2006 cooperative congressional election study. *Journal of Elections, Public Opinion and Parties*, 18, 355-366.

Wood, S.N. (2006). *Generalized Additive Models: An Introduction with R*, Chapman and Hall/CRC.

Yang, S., and Ding, P. (2020). Combining multiple observational data sources to estimate causal effects. *Journal of the American Statistical Association*, 115, 1540-1554.

Yang, S., and Kim, J.K. (2020). Statistical data integration in survey sampling: A review. *Japanese Journal of Statistics and Data Science*, 3, 625-650.

Yang, S., Kim, J.K. and Song, R. (2020). Doubly robust inference when combining probability and non-probability samples with high-dimensional data. *Journal of the Royal Statistical Society, Series B, (Statistical Methodology)*, 82, 445-465.

Yang, S., Zeng, D. and Wang, X. (2020). Improved inference for heterogeneous treatment effects using real-world data subject to hidden confounding, *arXiv preprint arXiv:2007.12922*.

Survey Methodology, June 2021
Vol. 47, No. 1, pp. 59-74
Statistics Canada, Catalogue No. 12-001-X

Sample empirical likelihood approach under complex survey design with scrambled responses

Sixia Chen, Yichuan Zhao and Yuke Wang[1]

Abstract

One effective way to conduct statistical disclosure control is to use scrambled responses. Scrambled responses can be generated by using a controlled random device. In this paper, we propose using the sample empirical likelihood approach to conduct statistical inference under complex survey design with scrambled responses. Specifically, we propose using a Wilk-type confidence interval for statistical inference. Our proposed method can be used as a general tool for inference with confidential public use survey data files. Asymptotic properties are derived, and the limited simulation study verifies the validity of theory. We further apply the proposed method to some real applications.

Key Words: Empirical likelihood; Scrambled responses; Statistical disclosure control; Survey data.

1. Introduction

The survey sampling technique has been shown to be one of the most effective ways to collect representative information for the underlying study population of interest; see Kish (1965) and Cochran (1977), among others. This approach has been used frequently in practice to obtain important information related to health, social economics, and public opinions. However, data collection by using a complex sampling design without careful control of statistical disclosure may lead to low response rate and large measurement error (Hundepool, Domingo-Ferrer, Franconi, Giessing, Nordholt, Spicer and Wolf, 2012). Statistical disclosure control (SDC) has been defined as one of few necessary steps to release public use files by agencies such as the US Census Bureau. For instance, Krenzke, Li, Freedman, Judkins, Hubble, Roisman and Larsen (2011) produced transportation data products from the Amercian Community Survey that comply with disclosure rules. Gouweleeuw, Kooiman, Willenborg and Wolf (1998) discussed statistical data protection at Statistics Netherlands.

The idea underlying SDC is to generate some perturbation based on the original raw data file so that the risk of identifying individuals is tiny and the utility of the perturbed data file is high. Currently, there are many SDC approaches including data coarsening, variable suppression, data swapping (Fienberg and McIntyre, 2005), Parametric model-based multivariate sequential replacement (Raghunathan, Lepkowski, van Hoewyk and Solenberger, 2001), and scrambled responses or randomized response methods (Horvitz, Shah and Simmons, 1967; Fox and Tracy, 1986). For more information about those approaches, see Hundepool et al. (2012).

Inference after SDC is an important and challenging problem. Statistical analysis without taking into account SDC leads to a biased variance estimation (Raghunathan, Reiter and Rubin, 2003). Raghunathan

1. Sixia Chen, Department of Biostatistics and Epidemiology, University of Oklahoma, Oklahoma City, OK 73104, U.S.A. E-mail: sixia-chen@ouhsc.edu; Yichuan Zhao, Department of Mathematics and Statistics, Georgia State University, Atlanta, GA 30303, U.S.A.; Yuke Wang, Department of Mathematics and Statistics, Georgia State University, Atlanta, GA 30303, U.S.A.

et al. (2003) proposed using the multiple imputation (MI) procedure to generate perturbed data files and using the Rubin's variance estimator formula for inference. However, most agencies only seek to produce one public use file, instead of many files and the validity of MI depends on the well-known congeniality condition of Meng (1994). This condition may not hold under informative sampling design (Kim and Yang, 2017). Compared with other approaches, the scrambled responses approach is very easy to implement and has good compromise of risk and utility. In addition, valid statistical inference can be developed for most complex sampling designs. Warner (1965) first proposed using a randomization device, such as a deck of cards, to estimate the proportion of sensitive characters, such as induced abortions, drug used, and so on. Tracy and Mangat (1996) contains a comprehensive review of randomized response methods. One effective randomized response method (Scrambled responses technique) is a multiplicative model considered by Eichhorn and Hayre (1983). Bar-Lev, Bobovitch and Boukai (2004) proposed an improved version of their model. Saha (2011) discussed an optional scrambled randomized response technique for practical surveys. More recently, Singh and Kim (2011) proposed using a pseudo empirical likelihood estimator with a simple random sampling without replacement (SRSWOR) design under this model. However, they only considered a point estimation under the SRSWOR design, and their proposed method may not work for other sampling designs, such as probability proportional to size design.

Empirical likelihood approach was proposed by Hartley and Rao (1968) and studied by Owen (1988, 2001) and Qin and Lawless (1994) under traditional statistical settings. Under complex survey settings, Wu and Rao (2006) considered pseudo empirical likelihood approach. Chen and Kim (2014) proposed population and sample empirical likelihood methods which are more efficient than pseudo empirical likelihood method with high entropy designs. Berger and Torres (2016), Berger (2018a, 2018b) extended the sample empirical likelihood approach in Chen and Kim (2014) to a more general setting. In this paper, we only consider single stage sampling designs, which include Poisson sampling and stratified probability proportional to size sampling designs. Our proposed approach can be generalized to multi-stage design by using the method discussed in Berger (2018b). In surveys with multi-stage design, one challenge is that we need to specify the conditions of inclusion probabilities and consider the correlation of observations within the same cluster in different stages. We also consider interval estimation by using the sample empirical likelihood method considered in Chen and Kim (2014). After estimating the scale factor consistently, the adjusted pseudo empirical likelihood ratio converges to a standard Chi-square distribution, which can be used to construct the confidence interval. External aggregated auxiliary information, such as population size by age, gender, and race, can be naturally incorporated into our proposed method to improve the efficiency of the proposed estimators. Our proposed method is practical and can be used in most public-use survey data files, such as those from the National Health and Nutrition Examination Survey (NHANES), National Health Interview Survey (NHIS), and Behavioral Risk Factor Surveillance System (BRFSS).

The paper is organized as follows. Basic notations, research questions, and the Hájek estimator are introduced in Section 2. Section 3 discusses the proposed sample empirical likelihood method. One

simulation study is presented in Section 4. We apply the proposed methods to 2015-2016 National Health Nutrition and Examination Survey (NHANES) data in Section 5. In Section 6, we conclude this paper. All technique details are contained in the Appendix.

2. Preliminaries

Suppose the finite population $\mathcal{F}_N = (X_i, Y_i, i = 1, \ldots, N)$ is generated from some unknown super-population model, where Y_i is a study variable and X_i is a covariate. For ease of presentation, given $\mathcal{F}_N$, a random sample A is assumed to be selected from a single stage unstratified sampling design. Let I_i be the sampling indicator for unit i such that $I_i = 1$ if unit i is selected and 0 otherwise. Denote the first-order and second-order inclusion probabilities as $\pi_i = E(I_i)$ and $\pi_{ij} = E(I_i I_j)$ for $i, j = 1, \ldots, N$. Then, the sampling weight can be written as $d_i = \pi_i^{-1}$ and sample size is $n = \sum_{i=1}^{N} I_i$. Suppose the parameter of interest is $\theta_N = N^{-1}\sum_{i=1}^{N} Y_i$. Due to confidentiality, we plan to use scrambled responses Z_i of Y_i such that $Z_i = Y_i S_i$ with probability $1 - p$ and $Z_i = Y_i$ with probability p, where $E(S_i) = a$ and $V(S_i) = b^2$ with p, a, and b^2 known. Bar-lev et al. (2004) and Singh and Kim (2011) considered similar models. Instead of observing Y_i directly, we only observe the scrambled responses Z_i in the data file. Hájek estimator discussed in Hájek (1971) and Fuller (2009) has been used frequently in survey data analysis. Under certain regularity conditions, one can show that the following Hájek (HJ) type estimator is consistent:

$$\hat{\theta}_{\mathrm{HJ}} = \frac{1}{\hat{N}} \sum_{i \in A} d_i Y_i^*, \tag{2.1}$$

where $Y_i^* = Z_i \{(1 - p)\, a + p\}^{-1}$ and $\hat{N} = \sum_{i \in A} d_i$ since $E(\hat{N}) = N$ and

$$E\left(\sum_{i \in A} d_i Y_i^*\right) = \sum_{i=1}^{N} \{E(Y_i^*)\} = \sum_{i=1}^{N} \left[\{E(Y_i S_i)\,(1 - p) + Y_i p\}\{(1 - p)\, a + p\}^{-1}\right] = \sum_{i=1}^{N} Y_i.$$

The asymptotic properties of $\hat{\theta}_{\mathrm{HJ}}$ are described in the following Theorem 1, and the sketched proof is contained in Appendix B.

Theorem 1. *Under the regularity conditions in Appendix A, $\hat{\theta}_{\mathrm{HJ}}$ has the following asymptotic expansion*

$$\hat{\theta}_{\mathrm{HJ}} = \theta_N + \frac{1}{N} \sum_{i \in A} d_i (Y_i^* - \theta_N) + o_p(n^{-1/2}), \tag{2.2}$$

and

$$V_{\mathrm{HJ}}^{-1/2} (\hat{\theta}_{\mathrm{HJ}} - \theta_N) \to^d N(0, 1), \tag{2.3}$$

as $n, N \to \infty$ with

$$V_{\mathrm{HJ}} = V_1 + V_2, \tag{2.4}$$

where

$$V_1 = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\left(Y_i-\theta_N\right)\left(Y_j-\theta_N\right),$$

and

$$V_2 = \frac{(1-p)\left\{p(a-1)^2+b^2\right\}}{\left\{(1-p)\,a+p\right\}^2}\,\frac{1}{N^2}\sum_{i=1}^{N}Y_i^2.$$

Note that V_1 is the design variability of Hájek estimator for population mean θ_N by using the true values and V_2 is the additional variability generated by using scrambled responses. According to Theorem 1, the consistent estimator of V_{HJ} can be written as

$$\hat{V}_{\mathrm{HJ}} = \frac{1}{\hat{N}^2}\sum_{i\in A}\sum_{j\in A}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}}\,\frac{Y_i^*-\hat{\theta}_{\mathrm{HJ}}}{\pi_i}\,\frac{Y_j^*-\hat{\theta}_{\mathrm{HJ}}}{\pi_j}+\frac{(1-p)\left\{b^2+p(a-1)^2\right\}}{\left(b^2+a^2\right)(1-p)+p}\,\frac{1}{\hat{N}^2}\sum_{i\in A}d_iY_i^{*2}.$$

When $n/N = o(1)$, the second term above can be safely ignored. Therefore, we can use a traditional design consistent estimator with transformed variable Y_i^*. In the next section, we will propose using the pseudo empirical likelihood method to construct both point estimator and confidence interval when we have aggregated auxiliary information.

3. Proposed method

Population-level aggregated information is often available through census or large surveys, such as the American Community Survey (ACS). For instance, we may know the national-level population counts by gender, race, educational level, or income level. Incorporating such information into estimation will often reduce the coverage error and improve the efficiency of the estimators. In this section, we propose using the sample empirical likelihood (SEL) approach proposed by Chen and Kim (2014) to conduct point and interval estimation simultaneously. Suppose a population mean $\bar{X}_N = N^{-1}\sum_{i=1}^{N}X_i$ is known through some external resources. Then, the SEL estimator can be obtained by maximizing the following sample empirical log-likelihood function

$$l_s = \sum_{i\in A}\log(w_i), \tag{3.1}$$

subject to constraints

$$\sum_{i\in A}w_i = 1, \quad \sum_{i\in A}w_i\pi_i^{-1}\left(X_i-\bar{X}_N\right)=0, \quad w_i\geq 0, \tag{3.2}$$

and

$$\sum_{i\in A}w_i\pi_i^{-1}\left(Y_i^*-\theta\right)=0. \tag{3.3}$$

By maximizing objective function (3.1) subject to constraints in (3.2), the SEL weight can be written as

$$\hat{w}_i = \frac{1}{n} \frac{1}{1 + \hat{\lambda}\pi_i^{-1}\left(X_i - \bar{X}_N\right)},$$

with $\hat{\lambda}$ as the Lagrange multiplier, and it can be obtained by solving the second constraint in (3.2). Then, according to (3.3), the SEL estimator of θ_N can be written as

$$\hat{\theta}_{\text{SEL}} = \frac{\sum_{i\in A} \hat{w}_i \pi_i^{-1} Y_i^*}{\sum_{i\in A} \hat{w}_i \pi_i^{-1}}.$$

The following Theorem 2 contains asymptotic properties of the proposed SEL estimator $\hat{\theta}_{\text{SEL}}$. The sketched proof is contained in Appendix C.

Theorem 2. *Under the regularity conditions in Appendix A, $\hat{\theta}_{\text{SEL}}$ has the following asymptotic expansion*

$$\hat{\theta}_{\text{SEL}} = \theta_N + \frac{1}{N}\sum_{i\in A} d_i\left(Y_i^* - \theta_N\right) - B\,\frac{1}{N}\sum_{i\in A} d_i\left(X_i - \bar{X}_N\right) + o_p\left(n^{-1/2}\right), \tag{3.4}$$

where

$$B = \left\{\frac{1}{N}\sum_{i=1}^{N} \pi_i^{-1}\left(Y_i - \theta_N\right)\left(X_i - \bar{X}_N\right)\right\}\left\{\frac{1}{N}\sum_{i=1}^{N} \pi_i^{-1}\left(X_i - \bar{X}_N\right)\left(X_i - \bar{X}_N\right)^{\top}\right\}^{-1}$$

and

$$V_{\text{SEL}}^{-1/2}\left(\hat{\theta}_{\text{SEL}} - \theta_N\right) \to^d N(0, 1),$$

as $n, N \to \infty$ with

$$V_{\text{SEL}} = V_1^* + V_2,$$

where V_2 is defined in Theorem 1 and

$$V_1^* = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j}\,\eta_i\eta_j,$$

with $\eta_i = Y_i - \theta_N - B\left(X_i - \bar{X}_N\right)$.

Note that V_1^* is the design variability of optimal regression estimator which is less than V_1 defined in Theorem 1. The optimal regression estimator has been discussed by Fuller and Isaki (1981), Montanari (1987), and Rao (1994). According to Theorem 2, the consistent estimator of V_{SEL} can be written as

$$\hat{V}_{\text{SEL}} = \frac{1}{\hat{N}^2}\sum_{i\in A}\sum_{j\in A} \frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}}\,\frac{\hat{\eta}_i}{\pi_i}\,\frac{\hat{\eta}_j}{\pi_j} + \frac{(1-p)\left\{b^2 + p(a-1)^2\right\}}{\left(b^2 + a^2\right)(1-p) + p}\,\frac{1}{\hat{N}^2}\sum_{i\in A} d_i Y_i^{*2},$$

where $\hat{\eta}_i = Y_i^* - \hat{\theta}_{\text{SEL}} - \hat{B}\left(X_i - \bar{X}_N\right)$ with

$$\hat{B} = \left\{\sum_{i\in A} d_i^2\left(Y_i^* - \hat{\theta}_{\text{SEL}}\right)\left(X_i - \bar{X}_N\right)\right\}\left\{\sum_{i\in A} d_i^2\left(X_i - \bar{X}_N\right)\left(X_i - \bar{X}_N\right)^{\top}\right\}^{-1}.$$

When $n/N = o(1)$, the second term of $\hat{V}_{\text{SEL}}$ can be ignored. Under the simple random sampling (SRS) design, it can be shown that $\hat{\theta}_{\text{SEL}}$ is asymptotically equivalent to the following well-known regression estimator

$$\hat{\theta}_{\text{REG}} = \frac{1}{\hat{N}} \sum_{i \in A} d_i Y_i^* - \hat{B}_R \left(\frac{1}{\hat{N}} \sum_{i \in A} d_i X_i - \bar{X}_N \right), \tag{3.5}$$

where

$$\hat{B}_R = \left\{ \sum_{i \in A} d_i \left(Y_i^* - \hat{\theta}_{\text{HJ}} \right) \left(X_i - \bar{X}_N \right) \right\} \left\{ \sum_{i \in A} d_i \left(X_i - \bar{X}_N \right) \left(X_i - \bar{X}_N \right)^{\top} \right\}^{-1}.$$

However, for general design, $\hat{\theta}_{\text{SEL}}$ is different from $\hat{\theta}_{\text{REG}}$. Under Poisson sampling design, it can be shown that $\hat{\theta}_{\text{SEL}}$ is more efficient than $\hat{\theta}_{\text{REG}}$. Theorem 1 and Theorem 2 can be used to construct a Wald-type confidence interval for θ_N. The following Theorem 3 can be used to construct a Wilk-type confidence interval. The sketched proof of Theorem 3 is contained in Appendix D.

Theorem 3. *Define* $R_n(\theta_N) = 2\left\{ l_s\left(\hat{\theta}_{\text{SEL}}\right) - l_s(\theta_N) \right\}$, *where* $l_s(\theta)$ *is defined in (3.1) with* w_i *satisfying (3.2) and (3.3). Then under the regularity conditions listed in Appendix A, as* $n, N \to \infty$, $c_1 c_2^{-1} R_n(\theta_N) \to^d \chi_1^2$, *where* $c_1 = N^{-2} \sum_{i=1}^{N} \pi_i^{-1} \eta_i^{*2}$ *with* $\eta_i^* = Y_i^* - \theta_N - B(X_i - \bar{X}_N)$ *and* $c_2 = V_{\text{SEL}}$.

The estimator of c_1 and c_2 can be written as

$$\hat{c}_1 = \hat{N}^{-2} \sum_{i \in A} \pi_i^{-2} \left\{ Y_i^* - \hat{\theta}_{\text{SEL}} - \hat{B}\left(X_i - \bar{X}_N\right) \right\}^2,$$

and $\hat{c}_2 = \hat{V}_{\text{SEL}}$. Theorem 3 can be used to construct a Wilk-type confidence interval for θ_N.

4. Simulation study

In the simulation study, we consider finite population (X_i, Y_i), $i = 1, 2, \ldots, N$ for $N = 10{,}000$. X_i is uniformly distributed over [0, 1] and $Y_i = m(X_i) + \varepsilon_i$ with $\varepsilon_i \sim N(0, 0.01)$. Four functions $m(x)$ are listed below:

(A). $m_1(x) = 2 + 2(x - 0.5)$,

(B). $m_2(x) = 2 + 2(x - 0.5)^2$,

(C). $m_3(x) = 2 + 2(x - 0.5) + \exp\left(-200(x - 0.5)^2\right)$,

(D). $m_4(x) = 2 + 2(x - 0.5)\Delta(x < 0.6) + 0.6\Delta(x \geq 0.6)$, where $\Delta(B)$ is the binary indicator function for condition B such that $\Delta(B) = 1$ if condition B is satisfied and 0 otherwise.

We generated $B = 5{,}000$ Monte Carlo samples from Poisson sampling with inclusion probabilities $\pi_i = nk_i \Big/ \sum_{j=1}^{N} k_j$, where the size variable $k_j = \max\left(0.5Y_j + 2, 1\right) + u_j$ with $u_j \sim \chi^2(1)$. We considered sample sizes $n = 40$, 50, 100 and 200. For each Monte Carlo sample, the scrambled responses Z_i were generated with $p = 0.6$, and $S_i \sim N(1.5, 0.2/1.5)$. Suppose we only observe X_i and Z_i in the

sample. The performance of the HJ estimator and the proposed SEL estimator were compared with the estimate population mean of Y, which is $\theta_0 = E(Y)$. The results are shown in Table 4.1.

We computed Monte Carlo bias $\text{MCB} = B^{-1}\sum_{b=1}^{B}\hat{\theta}_b - \theta_0$, Monte Carlo standard error $\text{MCSE} = \left\{B^{-1}\sum_{b=1}^{B}\left(\hat{\theta}_b - \bar{\theta}\right)^2\right\}^{1/2}$ with $\bar{\theta} = B^{-1}\sum_{b=1}^{B}\hat{\theta}_b$ and Monte Carlo mean squared error $\text{MCMSE} = \left\{B^{-1}\sum_{b=1}^{B}\left(\hat{\theta}_b - \theta_0\right)^2\right\}^{1/2}$. For variance estimation, we calculated coverage rate, average length of interval estimates, and percentage of relative bias of variance estimators $\text{RB} = 100 \times \left[\left(B^{-1}\sum_{b=1}^{B}\hat{V}_b\right)\left\{B^{-1}\sum_{b=1}^{B}\left(\hat{\theta}_b - \bar{\theta}\right)^2\right\}^{-1} - 1\right]$. Results obtained from the simulation are given in Table 4.1.

Table 4.1
Simulation results of Monte Carlo bias (MCB), Monte Carlo standard error (MCSE), and Monte Carlo mean squared error (MCMSE), coverage rate, average length of 95% confidence intervals, and relative bias (RB) for the Hájek (HJ) estimator and sample empirical likelihood (SEL) estimator

Setting		MCB		MCSE		MCMSE		Coverage Rate		Avg Length		RB	
Model	***n***	**HJ**	**SEL**	**HJ**	**SEL**	**HJ**	**SEL**	**HJ**	**SEL**	**HJ**	**SEL**	**HJ**	**SEL**
m_1	40	0.0035	0.0005	0.123	0.076	0.015	0.006	0.936	0.940	0.470	0.283	-0.027	-0.075
	50	0.0026	0.0006	0.110	0.069	0.012	0.005	0.939	0.941	0.420	0.255	-0.024	-0.078
	100	0.0009	0.0003	0.077	0.048	0.006	0.002	0.946	0.950	0.300	0.183	0.007	-0.000
	200	0.0006	-0.0002	0.054	0.033	0.003	0.001	0.944	0.954	0.211	0.130	-0.010	0.000
m_2	40	0.0006	0.0007	0.083	0.085	0.007	0.007	0.937	0.937	0.319	0.314	-0.020	-0.098
	50	-0.0004	-0.0008	0.074	0.075	0.005	0.006	0.939	0.944	0.286	0.283	-0.014	-0.066
	100	-0.0002	-0.0001	0.053	0.053	0.003	0.003	0.941	0.947	0.203	0.203	-0.036	-0.057
	200	-0.0007	-0.0006	0.037	0.037	0.001	0.001	0.945	0.949	0.144	0.144	0.002	-0.013
m_3	40	0.0022	0.0011	0.138	0.091	0.019	0.008	0.926	0.939	0.512	0.344	-0.081	-0.068
	50	0.0056	0.0028	0.119	0.081	0.014	0.007	0.941	0.942	0.460	0.312	-0.018	-0.045
	100	0.0011	0.0003	0.084	0.058	0.007	0.003	0.945	0.943	0.327	0.222	-0.011	-0.053
	200	-0.0002	-0.0006	0.059	0.041	0.003	0.002	0.950	0.952	0.230	0.157	-0.010	-0.028
m_4	40	0.0040	0.0012	0.119	0.080	0.014	0.006	0.938	0.937	0.460	0.296	-0.007	-0.089
	50	0.0008	0.0002	0.107	0.071	0.012	0.005	0.943	0.943	0.413	0.267	-0.020	-0.069
	100	0.0007	0.0006	0.075	0.049	0.006	0.002	0.942	0.945	0.293	0.190	-0.013	-0.036
	200	-0.0003	-0.0002	0.053	0.034	0.003	0.001	0.946	0.957	0.206	0.135	-0.018	0.029

For model m_1, m_3, and m_4, SEL has a smaller Monte Carlo bias, Monte Carlo standard error, and Monte Carlo mean squared error, especially for small sample sizes (n = 40 or 50). For model m_2, the two methods have comparable performance. For all four models, we found that, for most of the cases (14 of 16) the SEL estimators had a coverage rate higher than or equal to that of the HJ estimator, while the average length of confidence interval was shorter compared with the average length obtained with the HJ estimator. Both methods provided small relative biases of variance estimators. Overall, the proposed SEL outperformed HJ for most cases.

To test the sensitivity of the proposed approach, under current simulation study setups, we added noise, W_i, to the simulation. Then, $Y_i = m(\alpha X_i + (1-\alpha)W_i) + \varepsilon_i$ with α = 0, 0.1, 0.3, 0.5, 0.7, 0.9, 1, $X_i \sim \text{Uniform}(0, 1)$, $W_i \sim N(0, 1)$, and $\varepsilon_i \sim N(0, 0.01)$. Suppose we only observe X_i and Z_i (the scrambled response of Y_i) in the sample, the HJ estimator and SEL estimator were again compared. The results are shown in Tables 4.2 and 4.3. We found that as α decreases, the coverage rates of the SEL

estimator are smaller than those of the HJ estimator, and the average length of CI for SEL estimator is not shorter than that of the HJ estimator. Therefore, the SEL estimator has better performance than the HJ estimator, provided that most of the information is contained in the current covariate.

Table 4.2
Simulation results of the Hájek (HJ) estimator and sample empirical likelihood (SEL) estimator after adding noise

Setting		$\alpha = 0$				$\alpha = 0.1$				$\alpha = 0.3$			
		Coverage Rate		Avg Length		Coverage Rate		Avg Length		Coverage Rate		Avg Length	
Model	n	HJ	SEL	HJ	SEL	HJ	SEL	HJ	SEL	HJ	SEL	HJ	SEL
m_1	40	0.924	0.903	1.419	1.368	0.926	0.911	1.289	1.251	0.938	0.928	1.045	1.022
	50	0.926	0.915	1.292	1.256	0.928	0.920	1.146	1.125	0.937	0.930	0.958	0.938
	100	0.940	0.935	0.927	0.927	0.941	0.935	0.839	0.838	0.948	0.943	0.679	0.668
	200	0.949	0.941	0.651	0.657	0.942	0.943	0.589	0.593	0.948	0.948	0.478	0.469
m_2	40	0.942	0.943	1.872	1.909	0.929	0.930	1.328	1.358	0.933	0.933	1.455	1.458
	50	0.935	0.937	1.704	1.732	0.933	0.937	1.181	1.206	0.931	0.935	1.327	1.325
	100	0.941	0.947	1.191	1.202	0.942	0.949	0.843	0.854	0.945	0.948	0.931	0.927
	200	0.949	0.952	0.841	0.845	0.949	0.955	0.593	0.597	0.948	0.948	0.645	0.640
m_3	40	0.917	0.899	1.438	1.382	0.925	0.906	1.313	1.273	0.933	0.922	1.044	1.020
	50	0.922	0.908	1.297	1.264	0.928	0.916	1.154	1.131	0.939	0.935	0.927	0.911
	100	0.937	0.928	0.960	0.958	0.941	0.935	0.838	0.838	0.940	0.938	0.660	0.654
	200	0.940	0.940	0.674	0.679	0.945	0.944	0.615	0.619	0.945	0.941	0.474	0.467
m_4	40	0.903	0.885	1.226	1.167	0.912	0.894	0.994	0.947	0.927	0.909	0.518	0.511
	50	0.921	0.912	1.093	1.057	0.917	0.912	0.902	0.870	0.928	0.918	0.460	0.457
	100	0.931	0.925	0.805	0.802	0.936	0.935	0.646	0.644	0.935	0.931	0.337	0.338
	200	0.941	0.939	0.581	0.585	0.936	0.939	0.460	0.462	0.945	0.946	0.236	0.237

Table 4.3
Simulation results of the Hájek (HJ) estimator and sample empirical likelihood (SEL) estimator after adding noise

Setting		$\alpha = 0.5$				$\alpha = 0.7$				$\alpha = 0.9$			
		Coverage Rate		Avg Length		Coverage Rate		Avg Length		Coverage Rate		Avg Length	
Model	n	HJ	SEL	HJ	SEL	HJ	SEL	HJ	SEL	HJ	SEL	HJ	SEL
m_1	40	0.934	0.933	1.002	0.934	0.933	0.935	1.091	0.959	0.937	0.940	1.292	1.096
	50	0.935	0.936	0.902	0.841	0.939	0.935	0.979	0.862	0.936	0.938	1.156	0.986
	100	0.947	0.948	0.635	0.596	0.944	0.949	0.697	0.616	0.946	0.948	0.820	0.705
	200	0.951	0.949	0.451	0.421	0.947	0.945	0.493	0.437	0.951	0.951	0.579	0.500
m_2	40	0.933	0.936	2.371	2.139	0.938	0.934	3.418	2.469	0.933	0.942	5.095	2.980
	50	0.940	0.941	2.148	1.938	0.940	0.937	3.057	2.210	0.945	0.944	4.583	2.687
	100	0.939	0.941	1.493	1.345	0.948	0.946	2.196	1.588	0.948	0.951	3.223	1.916
	200	0.942	0.942	1.054	0.938	0.944	0.947	1.545	1.113	0.949	0.947	2.264	1.356
m_3	40	0.939	0.935	1.004	0.940	0.935	0.937	1.101	0.970	0.939	0.947	1.288	1.093
	50	0.937	0.935	0.890	0.832	0.938	0.942	0.978	0.864	0.936	0.940	1.152	0.982
	100	0.946	0.945	0.635	0.595	0.951	0.952	0.698	0.616	0.948	0.952	0.821	0.706
	200	0.949	0.950	0.450	0.420	0.943	0.948	0.493	0.437	0.952	0.952	0.579	0.500
m_4	40	0.937	0.942	0.365	0.358	0.936	0.941	0.362	0.354	0.932	0.938	0.362	0.354
	50	0.935	0.939	0.326	0.322	0.939	0.943	0.325	0.320	0.938	0.947	0.324	0.320
	100	0.941	0.948	0.232	0.230	0.948	0.953	0.230	0.229	0.941	0.946	0.231	0.229
	200	0.947	0.948	0.165	0.164	0.942	0.944	0.163	0.163	0.949	0.951	0.163	0.163

5. Real application

In this section, we applied the proposed approach to 2015-2016 National Health and Nutrition Examination Survey (NHANES) to evaluate its practical performance. NHANES provides timely health- and nutrition-related information for the noninstitutionalized civilian resident population of the United States. It uses a complex, multistage probability design based on in-person survey to collect information. (see https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/overview.aspx?BeginYear=2015 for more information). The sample size for the 2015-2016 NHANES is about 9,000. We treated the original NHANES sample as a finite population and selected one sample by using a simple random sampling design with sample sizes (n) as 30, 40, 50, 100, and 200, respectively. Suppose our parameters of interest include population means of systolic blood pressure, diastolic blood pressure, HDL cholesterol, and total cholesterol. We created scramble responses for these parameters by using $p = 0.6$, $a = 1.5$, and $b^2 = 0.2/1.5$. In addition, body mass index (BMI) was selected as a covariate in the estimation process, since BMI is correlated with those study variables.

We compared the performances of two approaches, HJ and SEL, in terms of point estimates and interval estimates (Table 5.1). Point estimates obtained by using both methods were similar, and they were close to finite population parameters (120.47, 66.17, 54.43, and 180.25 for systolic blood pressure, diastolic blood pressure, HDL cholesterol, and total cholesterol), especially for larger sample sizes (Table 5.1). For systolic blood pressure, diastolic blood pressure, and total cholesterol, intervals produced by SEL shifted slightly to the right compared with the results produced by HJ for small sample sizes. However, when sample sizes increased, the results from the two approaches were similar. For HDL cholesterol, the results are comparable. The results from this application verified the validity of the proposed SEL approach.

Table 5.1
Point estimates and 95% CI for estimating means of different outcomes using scrambled response outcome and BMI from the NHANES data

	Systolic Blood Pressure in mm Hg		Diastolic Blood Pressure in mm Hg		HDL Cholesterol in mg/dL		Total Cholesterol in mg/dL	
n	**HJ**	**SEL**	**HJ**	**SEL**	**HJ**	**SEL**	**HJ**	**SEL**
30	124.5	124.5	67.7	69.4	57.9	57.6	187.0	188.3
	(112.3, 136.8)	(113.5, 139.6)	(61.5, 73.8)	(63.9, 75.2)	(50.3, 65.5)	(50.8, 65.9)	(160.0, 214.0)	(166.6, 225.5)
40	125.6	125.5	70.2	70.2	52.0	51.2	178.7	178.1
	(115.4, 135.8)	(116.5, 136.1)	(64.6, 75.8)	(64.9, 76.1)	(48.0, 56.0)	(47.3, 55.8)	(160.6, 196.8)	(162.1, 199.0)
50	118.3	116.9	67.1	67.1	57.1	56.8	173.7	173.3
	(110.2, 126.4)	(109.0, 126.1)	(60.9, 73.3)	(61.4, 73.8)	(50.8, 63.4)	(51.3, 63.2)	(160.2, 187.1)	(161.2, 187.8)
100	120.8	120.5	70.0	69.7	52.3	52.4	173.1	172.8
	(115.1, 126.5)	(115.1, 126.3)	(65.9, 74.0)	(65.9, 73.6)	(48.9, 55.7)	(49.2, 55.9)	(163.5, 182.7)	(164.0, 183.2)
200	124.1	123.9	67.6	67.5	54.0	53.8	181.4	181.5
	(119.4, 128.9)	(119.4, 128.8)	(64.9, 70.3)	(64.8, 70.3)	(51.1, 56.8)	(51.3, 56.5)	(172.7, 190.1)	(173.3, 190.9)

6. Conclusions

In this paper, we proposed a sample empirical likelihood (SEL)-based approach using scrambled responses to protect the confidentiality of complex survey data. The proposed SEL approach is easy to implement in practice and can be used as a general tool for statistical disclosure control. The idea of our proposed approach is to replace the true values by some scrambled values through random device, then the existing sample empirical likelihood approach can be applied with scrambled values to obtain the point estimation. However, the variance estimation and confidence interval estimation are different from that by treating the scrambled values as true values since we need to incorporate the randomness due to random device in the statistical inference. Such theoretical properties have been investigated and verified through simulation study and real data application. The SEL outperforms traditional approaches, such as HJ, by improving coverage rates and reducing the coverage lengths of confidence intervals. Chen and Kim (2014) has compared Wald-type CI and Wilk-type CI in the simulation studies by using sample empirical likelihood method. In general, the Wilk-type confidence intervals show better coverage properties than the Wald-type confidence intervals in terms of coverage rates. We would expect similar results by using our proposed approaches here. In future research, we will extend the proposed approach to estimate more general parameters, such as population quantiles and distribution functions. The corresponding statistical computational tools, such as R package, will also be developed.

Acknowledgements

Dr. Sixia Chen was partially supported by the Oklahoma Shared Clinical and Translational Resources (U54GM104938) with an Institutional Development Award (IDeA) from National Institute of General Medical Sciences. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The research of Yichuan Zhao was supported by the National Security Agency (NSA) Grant (H98230-12-1-0209) and the National Science Foundation Grant (DMS-1613176).

Appendix

A. Regularity conditions

We present the regularity conditions needed for proving Theorem 1 to Theorem 3 as following:

(C1). $c_1 < \pi_i N n^{-1} < c_2$ for $i = 1, 2, \ldots, N$ with $0 < c_1 < c_2$.

(C2). $n^{1/2}\left(N^{-1}\sum_{i=1}^{N} I_i \pi_i^{-1} Y_i - N^{-1}\sum_{i=1}^{N} Y_i\right) \to^d N(0, V_1)$ as $n \to \infty$ and $N \to \infty$, where $V_1 = nN^{-2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\pi_{ij} - \pi_i\pi_j) d_i Y_i d_j Y_j$.

(C3). $n^{1/2}\left(N^{-1}\sum_{i=1}^{N} I_i \pi_i^{-1} X_i - \frac{1}{N}\sum_{i=1}^{N} X_i\right) \to^d N(0, V_2)$ as $n \to \infty$ and $N \to \infty$, where $V_2 = nN^{-2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\pi_{ij} - \pi_i\pi_j) d_i X_i d_j X_j^{\top}$.

(C4). $N^{-1}\sum_{i=1}^{N} |Y_i|^4$ and $N^{-1}\sum_{i=1}^{N} \| X_i \|^4$ are bounded.

(C5). $\max_{i\in A}|Y_i| = o_p(n^{1/2})$ and $\max_{i\in A}\| X_i \| = o_p(n^{1/2})$.

B. Sketched proof of Theorem 1

$\hat{\theta}_{\mathrm{HJ}}$ can be written as the solution of estimating equation $\hat{U}_{\mathrm{HJ}}(\theta) = 0,$ where

$$\hat{U}_{\mathrm{HJ}}(\theta) = \frac{1}{N}\sum_{i\in A} d_i\left(Y_i^* - \theta\right).$$

Under the assumptions that $\hat{U}_{\mathrm{HJ}}(\theta)$ converges to $U_{\mathrm{HJ}}(\theta) = N^{-1}\sum_{i=1}^{N}(Y_i - \theta)$ uniformly, $E\left(Y^2\right) < \infty,$ and because of $U_{\mathrm{HJ}}(\theta_N) = 0,$ it can be shown that $\hat{\theta}_{\mathrm{HJ}} \to^p \theta_N.$ By using a Taylor expansion,

$$0 = \hat{U}_{\mathrm{HJ}}\left(\hat{\theta}_{\mathrm{HJ}}\right) = \hat{U}_{\mathrm{HJ}}(\theta_N) + \frac{\partial \hat{U}_{\mathrm{HJ}}(\theta_N)}{\partial\theta}\left(\hat{\theta}_{\mathrm{HJ}} - \theta_N\right) + o_p\left(n^{-1/2}\right).$$

After some algebra, it can be shown that

$$\hat{\theta}_{\mathrm{HJ}} = \theta_N + \frac{1}{N}\sum_{i\in A} d_i\left(Y_i^* - \theta_N\right) + o_p\left(n^{-1/2}\right).$$

Because

$$E\left(Y_i^*\right) = Y_i, \quad V\left(Y_i^*\right) = \frac{(1-p)\left\{b^2 + p(a-1)^2\right\}}{\{(1-p)a + p\}^2} Y_i^2, \tag{B.1}$$

$$\begin{aligned} V\left\{\frac{1}{N}\sum_{i\in A} d_i\left(Y_i^* - \theta_N\right)\right\} &= E\left\{\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j}\left(Y_i^* - \theta_N\right)\left(Y_j^* - \theta_N\right)\right\} \\ &\quad + V\left\{\frac{1}{N}\sum_{i=1}^{N}\left(Y_i^* - \theta_N\right)\right\}. \end{aligned} \tag{B.2}$$

According to (B.1), (B.2), and after some algebra, we can show that

$$V\left\{\frac{1}{N}\sum_{i\in A} d_i\left(Y_i^* - \theta_N\right)\right\} = V_{\mathrm{HJ}},$$

where V_{HJ} is defined in equation (2.4). Under the regularity conditions in Fuller and Isaki (1981), the asymptotic normality can be derived.

C. Sketched proof of Theorem 2

Define

$$\hat{U}_1(\lambda) = \frac{1}{N}\sum_{i\in A}\frac{\pi_i^{-1}\left(X_i - \bar{X}_N\right)}{1 + \lambda\pi_i^{-1}\left(X_i - \bar{X}_N\right)}$$

and

$$\hat{U}_2(\lambda, \theta) = \frac{1}{N} \sum_{i \in A} \frac{\pi_i^{-1}(Y_i^* - \theta)}{1 + \lambda \pi_i^{-1}(X_i - \bar{X}_N)}.$$

Then, $\hat{\theta}_{\text{SEL}}$ and $\hat{\lambda}$ are the solutions of $\hat{U}_1(\lambda) = \hat{U}_2(\lambda, \theta) = 0$. By using techniques similar to those of Chen and Kim (2014), it can be shown that $\hat{\lambda} = O_p(n^{-1/2})$ and $\hat{\theta}_{\text{SEL}} \to^p \theta_N$. Then, by using Taylor expansion, we have

$$0 = \hat{U}_1(\hat{\lambda}) = \hat{U}_1(0) + \frac{\partial \hat{U}_1(0)}{\partial \lambda} \hat{\lambda} + o_p(n^{-1/2}), \tag{C.1}$$

and

$$0 = \hat{U}_2(\hat{\lambda}, \hat{\theta}_{\text{SEL}}) = \hat{U}_2(0, \theta_N) + \frac{\partial \hat{U}_2(0, \theta_N)}{\partial \theta}(\hat{\theta}_{\text{SEL}} - \theta_N) + \frac{\partial \hat{U}_2(0, \theta_N)}{\partial \lambda} \hat{\lambda} + o_p(n^{-1/2}). \tag{C.2}$$

According to (C.1), (C.2), and after some algebra, it can be shown that

$$\hat{\lambda} = \left\{ \frac{1}{N} \sum_{i=1}^{N} \pi_i^{-1}(X_i - \bar{X}_N)(X_i - \bar{X}_N)^{\top} \right\}^{-1} \frac{1}{N} \sum_{i \in A} d_i (X_i - \bar{X}_N) + o_p(n^{-1/2}) \tag{C.3}$$

and

$$\hat{\theta}_{\text{SEL}} - \theta_N = \frac{1}{N} \sum_{i \in A} d_i (Y_i^* - \theta_N) - B \frac{1}{N} \sum_{i \in A} d_i (X_i - \bar{X}_N) + o_p(n^{-1/2}),$$

where B is defined in Theorem 2. Because

$$\begin{aligned} V(\hat{\theta}_{\text{SEL}}) = V\left(\frac{1}{N} \sum_{i \in A} d_i \eta_i^* \right) + o(n^{-1}) &= V\left(\frac{1}{N} \sum_{i \in A} d_i \eta_i \right) + E\left\{ V\left(\frac{1}{N} \sum_{i \in A} d_i \eta_i^* \,\middle|\, A \right) \right\} \\ &= V_2 + \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \eta_i \eta_j + o(n^{-1}), \end{aligned}$$

where V_2 is defined in Theorem 1, η_i is defined in Theorem 2 and $\eta_i^* = Y_i^* - \theta_N - B(X_i - \bar{X}_N)$. After some algebra, we can show that

$$\hat{V}_{\text{SEL}} = V_{\text{SEL}} + o(n^{-1}),$$

with V_{SEL} defined in Theorem 2. Furthermore, under the regularity conditions in Fuller and Isaki (1981), we obtain the asymptotic normality.

D. Sketched proof of Theorem 3

Because $\hat{\lambda} = O_p(n^{-1/2})$ and by using a Taylor expansion of $\log(1 + x)$ at $x = \hat{\lambda} \pi_i^{-1}(X_i - \bar{X}_N)$ and (C.3), we have

$$
\begin{aligned}
l_s\left(\hat{\theta}_{\mathrm{SEL}}\right) &= \sum_{i\in A}\log\left\{\frac{1}{n}\,\frac{1}{1+\hat{\lambda}\pi_i^{-1}\left(X_i-\bar{X}_N\right)}\right\}\\
&= n\log\left(\frac{1}{n}\right)-\sum_{i\in A}\log\left\{1+\hat{\lambda}\pi_i^{-1}\left(X_i-\bar{X}_N\right)\right\}\\
&= n\log\left(\frac{1}{n}\right)-\sum_{i\in A}\left\{\hat{\lambda}^{\top}\pi_i^{-1}\left(X_i-\bar{X}_N\right)-\frac{1}{2}\hat{\lambda}^{\top}\pi_i^{-2}\left(X_i-\bar{X}_N\right)^{\otimes 2}\hat{\lambda}\right\}+o_p(1)\\
&= n\log\left(\frac{1}{n}\right)-\frac{N}{2}\,\frac{1}{N}\sum_{i\in A}\pi_i^{-1}\left(X_i-\bar{X}_N\right)^{\top}\left\{\frac{1}{N}\sum_{i=1}^{N}\pi_i^{-1}\left(X_i-\bar{X}_N\right)^{\otimes 2}\right\}^{-1}\\
&\quad\times\frac{1}{N}\sum_{i\in A}\pi_i^{-1}\left(X_i-\bar{X}_N\right)+o_p(1), \qquad \text{(D.1)}
\end{aligned}
$$

with $a^{\otimes 2}=aa^{\top}$. We now consider to maximize $l_s=\sum_{i\in A}\log(w_i)$, subject to the following constraints

$$
\sum_{i\in A}w_i=1,\qquad \sum_{i\in A}w_i\pi_i^{-1}\left(X_i-\bar{X}_N\right)=0, \qquad \text{(D.2)}
$$

and

$$
\sum_{i\in A}w_i\pi_i^{-1}\eta_i^{*}=0, \qquad \text{(D.3)}
$$

where $\eta_i^{*}=Y_i^{*}-\theta_N-B\left(X_i-\bar{X}_N\right)$. The above constraints are equivalent with the original constraints (3.2) and (3.3). Define $u_i=\left(X_i^{\top}-\bar{X}_N^{\top},\,\eta_i^{*}\right)^{\top}$. Therefore, by using a similar argument, we have

$$
\begin{aligned}
l_s(\theta_N) &= \sum_{i\in A}\log\left\{\frac{1}{n}\,\frac{1}{1+\hat{\lambda}(\theta_N)\pi_i^{-1}u_i}\right\}\\
&= n\log\left(\frac{1}{n}\right)-\sum_{i\in A}\log\left\{1+\hat{\lambda}(\theta_N)\pi_i^{-1}u_i\right\}\\
&= n\log\left(\frac{1}{n}\right)-\sum_{i\in A}\left\{\hat{\lambda}^{\top}(\theta_N)\pi_i^{-1}u_i-\frac{1}{2}\hat{\lambda}^{\top}(\theta_N)\pi_i^{-2}u_i^{\otimes 2}\hat{\lambda}(\theta_N)\right\}+o_p(1)\\
&= n\log\left(\frac{1}{n}\right)-\frac{N}{2}\,\frac{1}{N}\sum_{i\in A}\pi_i^{-1}u_i^{\top}\left\{\frac{1}{N}\sum_{i=1}^{N}\pi_i^{-1}u_i^{\otimes 2}\right\}^{-1}\times\frac{1}{N}\sum_{i\in A}\pi_i^{-1}u_i+o_p(1)\\
&= n\log\left(\frac{1}{n}\right)-\frac{N}{2}\,\frac{1}{N}\sum_{i\in A}\pi_i^{-1}\left(X_i-\bar{X}_N\right)^{\top}\left\{\frac{1}{N}\sum_{i=1}^{N}\pi_i^{-1}\left(X_i-\bar{X}_N\right)^{\otimes 2}\right\}^{-1}\\
&\quad\times\frac{1}{N}\sum_{i\in A}\pi_i^{-1}\left(X_i-\bar{X}_N\right)-\frac{N}{2}\,\frac{1}{N}\sum_{i\in A}\pi_i^{-1}\eta_i^{*\top}\left\{\frac{1}{N}\sum_{i=1}^{N}\pi_i^{-1}\eta_i^{*\otimes 2}\right\}^{-1}\times\frac{1}{N}\sum_{i\in A}\pi_i^{-1}\eta_i^{*}+o_p(1) \qquad \text{(D.4)}
\end{aligned}
$$

provided $\sum_{i=1}^{N}\pi_i^{-1}\left(X_i-\bar{X}_N\right)\eta_i=0$. According to (D.1), (D.4), and after some algebra, we have

$$
\begin{aligned}
\frac{c_1}{c_2}R_n(\theta_N) &= \frac{2c_1}{c_2}\left\{l_s\left(\hat{\theta}_{\mathrm{SEL}}\right)-l_s(\theta_N)\right\}=\frac{c_1}{c_2}\left\{\frac{1}{N}\sum_{i\in A}\pi_i^{-1}\eta_i^{*}\right\}^2\left(\frac{1}{N^2}\sum_{i=1}^{N}\pi_i^{-1}\eta_i^{*2}\right)^{-1}\\
&= \left\{V\left(\frac{1}{N}\sum_{i\in A}\pi_i^{-1}\eta_i^{*}\right)^{-1/2}\frac{1}{N}\sum_{i\in A}\pi_i^{-1}\eta_i^{*}\right\}^2\ \to^{d}\ \chi_1^2. \qquad \text{(D.5)}
\end{aligned}
$$

Therefore, Theorem 3 is proven.

References

Bar-Lev, S.K., Bobovitch, E. and Boukai, B. (2004). A note on randomized response models for quantitative data. *Metrika*, 60, 255-260.

Berger, Y.G. (2018a). Empirical likelihood approaches in survey sampling. *The Survey Statistician*, 78, 22-31.

Berger, Y.G. (2018b). An empirical likelihood approach under cluster sampling with missing observations. *Annals of the Institute of Statistical Mathematics*, doi:10.1007/s10463-018-0681-x.

Berger, Y.G., and Torres, O.D.L.R. (2016). An empirical likelihood approach for inference under complex sampling design. *Journal of the Royal Statistical Society, Series B*, 78(2), 319-341.

Chen, S., and Kim, J.K. (2014). Population empirical likelihood for nonparametric inference in survey sampling. *Statistica Sinica*, 24, 335-355.

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

Eichhorn, B.H., and Hayre, L.S. (1983). Scrambled randomized response methods for obtaining sensitive quantitative data. *Journal of Statistical Planning and Inference*, 7, 307-316.

Fienberg, S.E., and McIntyre, J. (2005). Data swapping: Variations on a theme by Dalenius and Reiss. *Journal of Official Statistics*, 21, 309-323.

Fox, J.A., and Tracy, P.E. (1986). *Randomized Response: A Method for Sensitive Surveys*. Beverly Hills, CA: Sage.

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

Fuller, W.A., and Isaki, C.T. (1981). Survey design under superpopulation models. In *Current Topics in Survey Sampling*, (Eds., D. Krewski, J.N.K. Rao, and R. Platek). New York: Academic Press, 199-226.

Gouweleeuw, J.M., Kooiman, P., Willenborg, L.C.R. and Wolf, P. (1998). Post randomization for statistical disclosure control: Theory and implementation. *Journal of Official Statistics*, 14, 463-478.

Hájek, J. (1971). Comment on "An essay on the logical foundations of survey sampling, Part one". In *The Foundations of Survey Sampling*, (Eds., V.P. Godambe and D.A. Sprott), Holt, Rinehart, and Winston, 236.

Hartley, H.O., and Rao, J.N.K. (1968). A new estimation theory for sample surveys. *Biometrika*, 55, 547-557.

Horvitz, D.G., Shah, B.V. and Simmons, W.R. (1967). The unrelated question randomized response model. In *Proceedings of the Social Statistics Section,* American Statistical Association, 65-72.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E., Spicer, K. and Wolf, P. (2012). *Statistical Disclosure Control*, Wiley Series In Survey Methodology.

Kim, J.K., and Yang, S. (2017). A note on multiple imputation under informative sampling. *Biometrika*, 104, 221-228.

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

Krenzke, T., Li, J., Freedman, M., Judkins, D., Hubble, D., Roisman, R. and Larsen, M. (2011). Producing transportation data products from the American Community Survey that comply with disclosure rules. Washington, DC: National Cooperative Highway Research Program, Transportation Research Board, National Academy of Sciences.

Meng, X.L. (1994). Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statistical Science*, 9, 538-573.

Montanari, G.E. (1987). Post-sampling efficient Q-R prediction in large-sample surveys. *International Statistical Review*, 55, 191-202.

Owen, A.B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75, 237-249.

Owen, A.B. (2001). *Empirical Likelihood*. New York: Chapman and Hall.

Qin, J., and Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics*, 22, 300-325.

Raghunathan, T.E., Reiter, J.P. and Rubin, D.B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, 19, 1-16.

Raghunathan, T.E., Lepkowski, J.M., van Hoewyk, J., and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27, 1, 85-95. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2001001/article/5857-eng.pdf.

Rao, J.N.K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics*, 10, 153-165.

Saha, A. (2011). An optional scrambled randomized response technique for practical surveys. *Metrika*, 73, 139-149.

Singh, S., and Kim, J.M. (2011). A pseudo-empirical log-likelihood estimator using scrambled responses. *Statistics and Probability Letters*, 81, 345-351.

Tracy, D.S., and Mangat, N.S. (1996). Some developments in randomized response sampling during the last decade-a follow up of review by Chaudhuri and Mukerjee. *Journal of Applied Statistical Science*, 4, 147-158.

Warner, S.L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60, 63-69.

Wu, C., and Rao, J.N.K. (2006). Pseudo-empirical likelihood ratio confidence intervals for complex surveys. *Canadian Journal of Statistics*, 34, 359-375.

Survey Methodology, June 2021
Vol. 47, No. 1, pp. 75-89
Statistics Canada, Catalogue No. 12-001-X

A method to find an efficient and robust sampling strategy under model uncertainty

Edgar Bueno and Dan Hedlin[1]

Abstract

We consider the problem of deciding on sampling strategy, in particular sampling design. We propose a risk measure, whose minimizing value guides the choice. The method makes use of a superpopulation model and takes into account uncertainty about its parameters through a prior distribution. The method is illustrated with a real dataset, yielding satisfactory results. As a baseline, we use the strategy that couples probability proportional-to-size sampling with the difference estimator, as it is known to be optimal when the superpopulation model is fully known. We show that, even under moderate misspecifications of the model, this strategy is not robust and can be outperformed by some alternatives.

Key Words: Sampling design; GREG estimator; Risk Measure.

1. Introduction

We consider the problem of choosing strategy, in particular the design, for the estimation of the total of a study variable in a finite population when a set of J auxiliary variables is available in a list sampling frame. We focus on the estimation of the total.

The decision about sampling strategy involves parameters which are unknown at the stage when the decision needs to be taken. After data collection the parameters can be estimated, although sometimes only under some assumptions. In practice, we often use data from previous waves of a repeated survey, frame variables or data from another survey that is similar to the one at planning stage. There is a risk that the available data do not give reliable information about relevant parameters. The method presented here involves a risk measure, which takes into account the possibility of being misled by inaccurate or incorrect beliefs about the values of the needed parameters. The risk measure is derived for the difference and the generalized regression estimators. Other than that, the measure is general. This measure and the discussion of its practical use are the main result of this paper.

One aim when selecting and devising the sampling strategy is efficiency in terms of small mean-squared error. The definition of "efficiency" is not unique, however, as it depends on the inference approach. Under the design-based approach, Godambe (1955), Lanke (1973) and Cassel, Särndal and Wretman (1977) show that there is no uniformly best linear estimator, in the sense of being best for all populations. There is no best design either. Therefore, a traditional approach for defining the strategy has been to assume that the finite population is a realization of some superpopulation model. The strategy is then defined in such a way that it minimizes the model expected value of the design mean-squared error, a parameter called anticipated mean-squared error. The adjective "anticipated" was first introduced by Isaki

1. Edgar Bueno, Department of Statistics, Stockholm University, 106 91 Stockholm, Sweden. E-mail: edgar.bueno@stat.su.se; Dan Hedlin, Department of Statistics, Stockholm University, 106 91 Stockholm, Sweden. E-mail: dan.hedlin@stat.su.se.

and Fuller (1982) to emphasize the fact that this is a conceptual mean-squared error which is calculated in advance to sampling, based only on information available prior to sampling.

Assuming that a superpopulation model holds and its parameters are known, several authors have shown that the optimal strategy should make use of a probability proportional-to-size sampling design (e.g., Hájek, 1959; Cassel, Särndal and Wretman, 1976; Nedyalkova and Tillé, 2008). In practice, however, there is not even a consensus about the existence of a generating model, let alone what model to rely on. And even if there is a model, its parameters are unknown. There is evidence, rather empirical, that probability proportional-to-size sampling is not robust towards model misspecifications (e.g., Holmberg and Swensson, 2001). A second result of this paper is to provide some theoretical evidence of this fact.

Many articles discuss robustness in the survey sampling field. Beaumont, Haziza and Ruiz-Gazen (2013), for instance, propose a robust estimator that downweights influential observations; Royall and Herson (1973) consider robustness under polynomial models; Bramati (2012) and Zhai and Wiens (2015) propose robust stratification methods. We provide theoretical evidence of lack of robustness of proportional-to-size sampling and propose a method for assisting in the decision about the sampling design.

The contents of the paper are arranged as follows. The optimal strategy under the superpopulation model is defined in Section 2. The lack of robustness of this strategy when the model is misspecified is studied in Section 3. The method for assisting on the choice of the sampling design is presented in Section 4. In Section 5, the risk measure introduced in the previous section is extended to be used together with the GREG estimator. Section 6 presents numerical illustrations of the results in the paper. First, we illustrate the lack of robustness of probability proportional-to-size sampling and the flexibility of the GREG estimator with a small simulation study. Second, we illustrate the implementation of the risk measure with real survey data. Finally, Section 7 presents some conclusions.

2. Optimal strategy under the superpopulation model

Let U be a finite population of size N with elements labeled $\{1, 2, \ldots, k, \ldots, N\}$. Let $x_k = (x_{1k}, x_{2k}, \ldots, x_{Jk})$ be a known vector of values of J auxiliary variables and y_k the unknown value of a study variable associated to unit $k \in U$. We are interested in the estimation of the total of y, $t_y = \sum_U y_k$.

Let Ω be the power set of U. A *sample* is any subset $s \in \Omega$ and a *sampling design* is a probability distribution on Ω, denoted by $P(S = s)$ or simply $p(s)$. Let $\pi_k = \sum_{s \ni k} p(s)$ be the *inclusion probability* of k and $\pi_{kl} = \sum_{s \supset \{k,l\}} p(s)$ the *joint inclusion probability* of k and l. A *probability sampling design* is a sampling design such that $\pi_k > 0$ for all $k \in U$.

An *estimator* is a real valued function of the sample, $\hat{t}_y = \hat{t}_y(S)$. By *strategy* we refer to the couple sampling design and estimator, $(p(\cdot), \hat{t}_y)$.

We consider only probability sampling designs with fixed sample size. As a convenient stepping stone we begin by considering unbiased linear estimators of the form

$$\hat{t}_y = \left(\sum_U z_k - \sum_s \frac{z_k}{\pi_k} \right) + \sum_s \frac{y_k}{\pi_k} = \sum_U z_k + \sum_s \frac{e_k}{\pi_k} \tag{2.1}$$

with z_k arbitrary known constants and $e_k = y_k - z_k$. This estimator is called the *difference estimator*. The estimator defined in this way is said to be *calibrated* on z as it satisfies $\hat{t}_z = \sum_U z_k$. Note that if $z_k = 0$ for all $k \in U$ the estimator reduces to $\hat{t}_y = \sum_s y_k / \pi_k$, that is, the Horvitz-Thompson estimator (Horvitz and Thompson, 1952). In later sections we focus on the generalized regression estimator (GREG).

The design Mean Squared Error (MSE) of the difference estimator is

$$\mathrm{MSE}_{\mathrm{p}}\left(\hat{t}_y\right) = \mathrm{MSE}_{\mathrm{p}}\left(\sum_s \frac{e_k}{\pi_k} \right) = \sum_U \sum_U (\pi_{kl} - \pi_k \pi_l) \frac{e_k}{\pi_k} \frac{e_l}{\pi_l}. \tag{2.2}$$

As mentioned in the introduction, due to the non-existence of an optimal strategy under the design-based approach, often a superpopulation model, ξ_0, is proposed and we search for an optimal strategy with respect to the *anticipated mean-squared error*,

$$\mathrm{MSE}_{\xi_0 \mathrm{p}}\left(\hat{t}_y\right) = \mathrm{E}_{\xi_0} \mathrm{MSE}_{\mathrm{p}}\left(\hat{t}_y\right) = \mathrm{E}_{\xi_0} \mathrm{E}_p \left(\left(\hat{t}_y - t_y \right)^2 \right). \tag{2.3}$$

We may assume that the y-values are realizations of the following model, denoted ξ_0,

$$Y_k = f(x_k \mid \delta_1) + \varepsilon_k$$

with

$$\mathrm{E}_{\xi_0}(\varepsilon_k) = 0, \quad \mathrm{V}_{\xi_0}(\varepsilon_k) = \sigma_0^2 g(x_k \mid \delta_2)^2 \quad \text{and} \quad \mathrm{E}_{\xi_0}(\varepsilon_k \varepsilon_l) = 0 \ \forall \ k \neq l \tag{2.4}$$

where $\delta = (\delta_1, \delta_2)$ is a vector of parameters, $f: \mathcal{R}^J \rightarrow \mathcal{R}$ and $g: \mathcal{R}^J \rightarrow \mathcal{R}^+$. The random sample s and the errors ε_k are assumed to be independent. Following Rosén (2000), the terms $f(x_k \mid \delta_1)$ and $g(x_k \mid \delta_2) > 0$ will be called *trend* and *spread*, respectively. The term trend should not in general be understood in a temporal sense, rather it refers to the development of y-values with x.

Note that under ξ_0, e_k in the difference estimator (2.1) is a random variable that represents the distance between the value of the study variable and z_k, i.e., $e_k = f(x_k \mid \delta_1) + \varepsilon_k - z_k$. Therefore $\mathrm{E}_{\xi_0} e_k = f(x_k \mid \delta_1) - z_k$ and $\mathrm{E}_{\xi_0} e_k^2 = \left(f(x_k \mid \delta_1) - z_k \right)^2 + \sigma_0^2 g(x_k \mid \delta_2)^2$. With some algebra, it can be seen from (2.2) and (2.3) that the anticipated MSE of the difference estimator becomes

$$\mathrm{MSE}_{\xi_0 \mathrm{p}}\left(\hat{t}_y\right) = \mathrm{MSE}_{\mathrm{p}}\left(\sum_s \frac{f(x_k \mid \delta_1) - z_k}{\pi_k} \right) + \sigma_0^2 \sum_U \left(\frac{1}{\pi_k} - 1 \right) g(x_k \mid \delta_2)^2 \tag{2.5}$$

Nedyalkova and Tillé (2008) derive the anticipated MSE in a more general case.

Tillé and Wilhelm (2017) give the anticipated MSE of the Horvitz-Thompson estimator. The second term in (2.5) is the Godambe-Joshi lower bound (e.g., Särndal, Swensson and Wretman, 1992, page 453).

The anticipated MSE in (2.5) is the sum of two positive terms. It is easy to see that if

1. the estimator is calibrated on $z_k = f(x_k|\delta_1)$ the first term vanishes and the anticipated MSE equals the Godambe-Joshi lower bound

$$\mathrm{MSE}_{\xi_0 \mathrm{p}}(\hat{t}_y) = \sigma_0^2 \sum_U \left(\frac{1}{\pi_k} - 1 \right) g(x_k|\delta_2)^2. \tag{2.6}$$

Furthermore, after imposing the fixed sample size restriction $\sum_U \pi_k = n$, if

2. the design is such that $\pi_k \propto g(x_k|\delta_2)$, denoted $\pi\mathrm{ps}(\delta_2)$, the second term is minimized and we obtain

$$\mathrm{MSE}^{\mathrm{opt}}_{\xi_0 \mathrm{p}}(\hat{t}_y) = \sigma_0^2 \left(\frac{1}{n} \left(\sum_U g(x_k|\delta_2) \right)^2 - \sum_U g(x_k|\delta_2)^2 \right).$$

Conditions 1 and 2 suggest the specific roles of the design and the estimator in the sampling strategy. The estimator should "explain" the trend in the calibration sense of condition 1. The design should "explain" the spread. A strategy that satisfies conditions 1 and 2 simultaneously will be called optimal. In the same sense, any estimator and any design satisfying, respectively, condition 1 and 2, will be called optimal. As this strategy plays an important role in this paper, we will denote it by $\pi\mathrm{ps}(\delta_2) - \mathrm{diff}(\delta_1)$.

3. Robustness under a misspecified model

If the finite population is a realization of the superpopulation model (2.4), and if f, g and δ were known, then an optimal strategy could be defined. In this section we study the robustness of this strategy when the model is misspecified.

We begin by defining how "misspecification" shall be understood in this paper. The *working model* ξ_0 reflects the beliefs the statistician has about the relation between the auxiliary variables x and the study variable y at the design stage. We shall assume that a true, unknown model ξ exists. Any deviation of ξ_0 with respect to ξ is a misspecification of the model. In order to keep the analysis tractable, we limit ourselves to the situation when the working model is of the form (2.4) and the true model, ξ, is

$$Y_k = f(x_k|\beta_1) + \varepsilon_k$$

with

$$\mathrm{E}_\xi(\varepsilon_k) = 0, \quad \mathrm{V}_\xi(\varepsilon_k) = \sigma^2 g(x_k|\beta_2)^2 \quad \text{and} \quad \mathrm{E}_\xi(\varepsilon_k \varepsilon_l) = 0 \ \forall\ k \neq l \tag{3.1}$$

where $\beta = (\beta_1, \beta_2)$ is a vector of parameters, f and g as in (2.4) and $\beta \neq \delta$. The random sample s and the errors ε_k are assumed to be independent.

Result 1. If ξ_0 is assumed when ξ is the true superpopulation model, the model expected value of the design MSE in (2.2), under the difference estimator satisfying condition 1 above, becomes

$$\mathrm{MSE}_{\xi\mathrm{p}}\left(\hat{t}_y\right) = \mathrm{MSE}_{\mathrm{p}}\left(\sum_s \frac{f\left(x_k \mid \beta_1\right) - f\left(x_k \mid \delta_1\right)}{\pi_k}\right) + \sigma^2 \sum_U \left(\frac{1}{\pi_k} - 1\right) g\left(x_k \mid \beta_2\right)^2. \qquad (3.2)$$

The result is proven by noting that $f\left(x_k \mid \delta_1\right)$ takes the role of z_k in (2.5) and by taking into account that $e_k = f\left(x_k \mid \beta_1\right) - f\left(x_k \mid \delta_1\right) + \varepsilon_k$, therefore $\mathrm{E}_\xi e_k = f\left(x_k \mid \beta_1\right) - f\left(x_k \mid \delta_1\right)$ and $\mathrm{E}_\xi e_k^2 = \left(f\left(x_k \mid \beta_1\right) - f\left(x_k \mid \delta_1\right)\right)^2 + \sigma^2 g\left(x_k \mid \beta_2\right)^2$. As the model is misspecified, we have deliberately avoided the use of the adjective "anticipated" in Result 1.

Using Result 1, it can be seen that for a design that satisfies condition 2 we obtain

$$\begin{aligned}\mathrm{MSE}_{\xi,\pi\mathrm{ps}}\left(\hat{t}_y\right) = {} & \left(\frac{\sum_U g\left(x_k \mid \delta_2\right)}{n}\right)^2 \mathrm{MSE}_{\pi\mathrm{ps}}\left(\sum_s \frac{f\left(x_k \mid \beta_1\right) - f\left(x_k \mid \delta_1\right)}{g\left(x_k \mid \delta_2\right)}\right) \\ & + \sigma^2 \sum_U \left(\frac{n\sum_U g\left(x_k \mid \delta_2\right)}{g\left(x_k \mid \delta_2\right)} - 1\right) g\left(x_k \mid \beta_2\right)^2. \qquad (3.3)\end{aligned}$$

It is now possible to see that, even under a mild misspecification as the one considered here, the strategy $\pi\mathrm{ps}\left(\delta_2\right) - \mathrm{diff}\left(\delta_1\right)$ is not optimal anymore, as its MSE (3.3) can be greater than the MSE obtained under other designs (3.2). In particular, the strategy using the correct model, i.e., $z_k = f\left(x_k \mid \beta_1\right)$ into the estimator and a design such that $\pi_k \propto g\left(x_k \mid \beta_2\right)$, would be more efficient than $\pi\mathrm{ps}\left(\delta_2\right) - \mathrm{diff}\left(\delta_1\right)$.

4. Guiding the choice of sampling design with the help of a risk measure

We have seen in Section 3 that even a simple misspecification of the working model might result in the strategy $\pi\mathrm{ps}\left(\delta_2\right) - \mathrm{diff}\left(\delta_1\right)$ not being optimal. It is therefore risky to accept a given model as correct without any type of assessment. While most of the information needed for an "objective" evaluation of the model is not available at the design stage, it is possible to reach some degree of confidence about the parameters in the working model that allows for comparing a set of designs and make the decision about which one to implement. In this section we propose a method to assist in the choice of the sampling design.

The model expected MSE (3.2) in Result 1 can be viewed as a function of β and σ^2, as everything else is available at the design stage. To begin with, let us assume that σ^2 is also known. Then we can write

$$L_{\mathrm{p}}\left(\beta\right) = \mathrm{MSE}_{\xi\mathrm{p}}\left(\beta \mid x, \delta, \sigma\right) = \mathrm{MSE}_{\mathrm{p}}\left(\sum_s \frac{f\left(x_k \mid \beta_1\right) - f\left(x_k \mid \delta_1\right)}{\pi_k}\right) + \sigma^2 \sum_U \left(\frac{1}{\pi_k} - 1\right) g\left(x_k \mid \beta_2\right)^2.$$

For any design, $p(\cdot)$, this function can be evaluated at any β and it indicates the loss incurred by assuming that δ is the true parameter when it is, in fact, β. We can assume a prior distribution on β, $h(\beta)$, and calculate the risk under h,

$$R(\mathrm{p}) \;=\; \mathrm{E}_h\left(\mathrm{MSE}_{\xi\mathrm{p}}(\beta \mid x,\, \delta,\, \sigma)\right) = \int_{\Theta} h(\beta) \;\cdot\; \mathrm{MSE}_{\xi\mathrm{p}}(\beta \mid x,\, \delta,\, \sigma)\, d\beta, \tag{4.1}$$

where Θ is the sample space of β. The design that yields the smallest risk should be chosen. Note that numerical integration methods (e.g., Monte Carlo simulation methods) may be needed to evaluate the risk (4.1).

In practice, σ^2 is unknown. We propose two ways for dealing with it. The first one is to see now the loss as a function of β and σ and calculate the risk as above, assuming a prior on the pair β and σ. The second one is to provide some "guess" about its value. This approach can use the fact that (Proof in the Appendix)

$$\sigma^2 \;\approx\; \frac{S_{f,f}}{\overline{g}^{\,2}}\left(\frac{1}{R_{f,y}^2} - 1\right) \tag{4.2}$$

where $S_{f,f} = \sum_U \left(f(x_k \mid \beta_1) - \overline{f}\right)^2 \big/ N$, $\overline{f} = \sum_U f(x_k \mid \beta_1)/N$, $\overline{g}^{\,2} = \sum_U g(x_k \mid \beta_2)^2 \big/ N$ and $R_{f,y}$ is the correlation between $f(x \mid \beta_1)$ and y. (In Example 3 below, we give a more convenient expression in a special case.) Although $R_{f,y}$ is unknown, for repeated surveys we do have some previous knowledge about it. In other cases it is often possible to have some reasonable "guess" about it.

It remains to comment on the choice of the prior distribution $h(\beta)$. The choice of the distribution and its parameters is subjective and defined by the statistician. Nevertheless, it should reflect the available knowledge about the model parameter β. In particular, $h(\beta)$ should be centered around $\beta = \delta$. Its variance should reflect how confident we are about the working model. Note that a full confidence on the working model would be a density with all its mass at $\beta = \delta$, in which case the risk (4.1) would be minimized by the πps design given by condition 2 in Section 2.

It might be argued that by introducing $h(\beta)$ an additional source of subjectivity has been added to the choice of the sampling design. The prior may add a certain Bayesian flavor to the process, but note that $h(\beta)$ is only needed for choosing the design. Hence, the inference is still design-based. Furthermore, relying on an assumed model is also subjective in choice of assumption and it does involve a risk. The risk measure in (4.1) allows for quantification of that risk.

5. The risk measure under the Generalized Regression Estimator

The difference estimator (2.1) requires that δ_1 is fully specified in order to calculate $f(x_k \mid \delta_1)$, which is undesirable from a practical standpoint. The generalized regression (GREG) estimator is an alternative that allows for the estimation of all or some of the components of δ_1 at the cost of introducing a small bias. In this section we adapt the material in Sections 2 to 4 to strategies using the GREG estimator.

We define the generalized regression estimator in a more general way than in Särndal et al. (1992) as follows. Let a_k $(k = 1, \ldots, N)$ be a weight defined by the statistician and $\delta_1 = (\delta_1^*, \delta_1^{**})$ where δ_1^* is fixed and δ_1^{**} is to be estimated. Let also

$$\hat{\delta}_{1s}^{**} = \text{argmin}_{\delta_1^{**}} \sum_s \frac{(y_k - f(x_k | \delta_1))^2}{a_k \pi_k}$$

and $\hat{\delta}_{1s} = (\delta_1^*, \hat{\delta}_{1s}^{**})$. The GREG estimator is

$$\hat{t}_{\text{greg}} = \left(\sum_U f(x_k | \hat{\delta}_{1s}) - \sum_s \frac{f(x_k | \hat{\delta}_{1s})}{\pi_k} \right) + \sum_s \frac{y_k}{\pi_k}. \tag{5.1}$$

An approximation to the design MSE of the GREG estimator is of the form (2.2) with $e_k = y_k - f(x_k | \hat{\delta}_{1U})$ where $\hat{\delta}_{1U} = (\delta_1^*, \hat{\delta}_{1U}^{**})$ and

$$\hat{\delta}_{1U}^{**} = \text{argmin}_{\delta_1^{**}} \sum_U \frac{(y_k - f(x_k | \delta_1))^2}{a_k}.$$

Example 1. Let us consider the case where $f(x_k | \delta_1) = \delta_{1,1} x_{1k}^{\delta_{1,J+1}} + \delta_{1,2} x_{2k}^{\delta_{1,J+2}} + \ldots + \delta_{1,J} x_{Jk}^{\delta_{1,2J}}$. Let $\delta_1^* = (\delta_{1,J+1}, \ldots, \delta_{1,2J})$, $\delta_1^{**} = (\delta_{1,1}, \ldots, \delta_{1,J})'$ and $x_k^{\delta} = (x_{1k}^{\delta_{1,J+1}}, \ldots, x_{Jk}^{\delta_{1,2J}})$. In this case we obtain

$$\hat{\delta}_{1s}^{**} = \left(\sum_s \frac{x_k^{\delta'} x_k^{\delta}}{a_k \pi_k} \right)^{-1} \sum_s \frac{x_k^{\delta'} y_k}{a_k \pi_k} \quad \text{and} \quad \hat{\delta}_{1U}^{**} = \left(\sum_U \frac{x_k^{\delta'} x_k^{\delta}}{a_k} \right)^{-1} \sum_U \frac{x_k^{\delta'} y_k}{a_k}.$$

Letting the exponents $\delta_1^* = (\delta_{1,J+1}, \ldots, \delta_{1,2J}) = (1, \ldots, 1)$, we obtain the classical expression of the GREG estimator found in Särndal et al. (1992).

Example 2. The case with only one auxiliary variable, i.e., $f(x_k | \delta_1) = \delta_{10} + \delta_{11} x_k^{\delta_{12}}$ with $a_k = 1$, $\delta_1^* = \delta_{12}$ and $\delta_1^{**} = (\delta_{10}, \delta_{11})'$ is known as the regression estimator. In this case we obtain the well known result that the design MSE can be approximated by expression (2.2) with $e_k = y_k - f(x_k | \hat{\delta}_{1U})$ where $f(x_k | \hat{\delta}_{1U}) = \hat{\delta}_{10} + \hat{\delta}_{11} x_k^{\delta_{12}}$ and

$$\hat{\delta}_{11} = \frac{N \sum_U x_k^{\delta_{12}} y_k - \sum_U x_k^{\delta_{12}} \sum_U y_k}{N \sum_U x_k^{2\delta_{12}} - \left(\sum_U x_k^{\delta_{12}} \right)^2} \quad \text{and} \quad \hat{\delta}_{10} = \frac{1}{N} \sum_U y_k - \hat{\delta}_{11} \frac{1}{N} \sum_U x_k^{\delta_{12}}.$$

The misspecified model

Let us consider again the situation where the statistician uses the working model (2.4) but the true model is of the form (3.1) with $\beta_1 = (\beta_1^*, \beta_1^{**})$, where β_1^* is the counterpart of the fixed component δ_1^*. The following result states a condition under which Result 1 is valid for the GREG estimator.

Result 2. If ξ_0 is assumed when ξ is the true superpopulation model, $\hat{\delta}_{1s}^{**} \to \hat{\delta}_{1U}^{**}$ as $n \to \infty$ and $\hat{\delta}_{1U}^{**}$ converges to some δ_1^{**} as $N \to \infty$, then

$$\mathrm{MSE}_{\xi \mathrm{p}}\left(\hat{t}_{\mathrm{greg}}\right) \rightarrow \mathrm{MSE}_{\mathrm{p}}\left(\sum_{s} \frac{f\left(x_k \mid \beta_1\right)-f\left(x_k \mid \delta_1\right)}{\pi_k}\right)+\sigma^2 \sum_{U}\left(\frac{1}{\pi_k}-1\right) g\left(x_k \mid \beta_2\right)^2 \qquad (5.2)$$

where $\delta_1 = \left(\delta_1^*, \delta_1^{**}\right)$.

Proof. Note that if $\hat{\delta}_{1s}^{**} \rightarrow \delta_{1U}^{**}$, then $\hat{\delta}_{1s} = \left(\delta_1^*, \hat{\delta}_{1s}^{**}\right) \rightarrow \left(\delta_1^*, \hat{\delta}_{1U}^{**}\right) = \hat{\delta}_{1U}$. Thus, $f\left(x_k \mid \hat{\delta}_{1s}\right) \rightarrow f\left(x_k \mid \hat{\delta}_{1U}\right)$. In turn, if $\hat{\delta}_{1U}^{**} \rightarrow \delta_1^{**}$, then $\hat{\delta}_{1U} = \left(\delta_1^*, \hat{\delta}_{1U}^{**}\right) \rightarrow \left(\delta_1^*, \delta_1^{**}\right) = \delta_1$. Thus $f\left(x_k \mid \hat{\delta}_{1U}\right) \rightarrow f\left(x_k \mid \delta_1\right)$. Therefore,

$$\begin{aligned}\mathrm{MSE}_{\xi \mathrm{p}}\left(\hat{t}_{\mathrm{greg}}\right) &= \mathrm{MSE}_{\xi \mathrm{p}}\left(\left(\sum_{U} f\left(x_k \mid \hat{\delta}_{1s}\right)-\sum_{s} \frac{f\left(x_k \mid \hat{\delta}_{1s}\right)}{\pi_k}\right)+\sum_{s} \frac{y_k}{\pi_k}\right) \\ &\rightarrow \mathrm{MSE}_{\xi \mathrm{p}}\left(\left(\sum_{U} f\left(x_k \mid \delta_1\right)-\sum_{s} \frac{f\left(x_k \mid \delta_1\right)}{\pi_k}\right)+\sum_{s} \frac{y_k}{\pi_k}\right),\end{aligned}$$

which, by Result 1, is (5.2).

Example 3 (Continuation of Example 1). Let the working model be as in Example 1 and the true model be $f\left(x_k \mid \beta_1\right) = \beta_{1,1} x_{1k}^{\beta_{1, J+1}} + \beta_{1,2} x_{2k}^{\beta_{1, J+2}} + \ldots + \beta_{1, J} x_{Jk}^{\beta_{1,2J}}$. Let also $\beta_1^* = \left(\beta_{1, J+1}, \ldots, \beta_{1,2J}\right)$, $\beta_1^{**} = \left(\beta_{1,1}, \ldots, \beta_{1, J}\right)'$ and $x_k^{\beta} = \left(x_{1k}^{\beta_{1, J+1}}, \ldots, x_{Jk}^{\beta_{1,2J}}\right)$. In this case, $\hat{\delta}_{1U}^{**} \rightarrow A\beta_1^{**}$, where

$$A = \left(\sum_{U} \frac{x_k^{\delta'} x_k^{\delta}}{a_k}\right)^{-1} \sum_{U} \frac{x_k^{\delta'} x_k^{\beta}}{a_k},$$

and (5.2) becomes

$$\mathrm{MSE}_{\xi \mathrm{p}}\left(\hat{t}_{\mathrm{greg}}\right) \rightarrow \mathrm{MSE}_{\mathrm{p}}\left(\sum_{s} \frac{\left(x_k^{\beta}-x_k^{\delta} A\right) \beta_1^{**}}{\pi_k}\right)+\sigma^2\left(\sum_{U} \frac{g\left(x_k \mid \beta_2\right)^2}{\pi_k}-\sum_{U} g\left(x_k \mid \beta_2\right)^2\right). \qquad (5.3)$$

Example 4 (Continuation of Example 2). Let the working model be as in Example 2 and the true model be $f\left(x_k \mid \beta_1\right) = \beta_{10} + \beta_{11} x_k^{\beta_{12}}$ with $\beta_1^* = \beta_{12}$ and $\beta_1^{**} = \left(\beta_{10}, \beta_{11}\right)'$. It can be shown that (5.2) becomes

$$\mathrm{MSE}_{\xi \mathrm{p}}\left(\hat{t}_{\mathrm{greg}}\right) \rightarrow \beta_{11}^2 \mathrm{MSE}_{\mathrm{p}}\left(\sum_{s} \frac{v_k}{\pi_k}\right)+\sigma^2\left(\sum_{U} \frac{g\left(x_k \mid \beta_2\right)^2}{\pi_k}-\sum_{U} g\left(x_k \mid \beta_2\right)^2\right) \qquad (5.4)$$

with

$$v_k = \left(x_k^{\beta_{12}}-\bar{x}^{\beta_{12}}\right)-\left(x_k^{\delta_{12}}-\bar{x}^{\delta_{12}}\right) \frac{S_{\beta,\delta}}{S_{\delta,\delta}}, \qquad (5.5)$$

and

$$\bar{x}^{\beta_{12}} = \frac{1}{N} \sum_{U} x_k^{\beta_{12}} \qquad S_{\beta,\delta} = \frac{1}{N-1} \sum_{U}\left(x_k^{\beta_{12}}-\bar{x}^{\beta_{12}}\right)\left(x_k^{\delta_{12}}-\bar{x}^{\delta_{12}}\right)$$

$$\bar{x}^{\delta_{12}} = \frac{1}{N} \sum_{U} x_k^{\delta_{12}} \qquad S_{\delta,\delta} = \frac{1}{N-1} \sum_{U}\left(x_k^{\delta_{12}}-\bar{x}^{\delta_{12}}\right)^2.$$

Note that (5.4) does not depend on β_{10}.

For the particular case developed in Examples 2 and 4, where $f(x_k \mid \beta) = \beta_{10} + \beta_{11} x_k^{\beta_{12}}$ and $f(x_k \mid \delta) = \delta_{10} + \delta_{11} x_k^{\delta_{12}}$, an alternative approximation of σ^2 is (Proof in the Appendix)

$$\sigma^2 \approx \beta_{11}^2 F_0 \quad \text{with} \quad F_0 = \frac{1}{\bar{x}^{2\beta_2}} \frac{S_{1,\beta}^2}{S_{1,1}} \left(\frac{1}{R_{x,y}^2} - \frac{1}{R_{1,\beta}^2} \right) \tag{5.6}$$

where

$$\bar{x}^{2\beta_2} = \frac{1}{N} \sum_U x_k^{2\beta_2} \qquad S_{1,\beta} = \frac{1}{N} \sum_U (x_k - \bar{x})(x_k^{\beta_{12}} - \bar{x}^{\beta_{12}}) \qquad S_{1,1} = \frac{1}{N} \sum_U (x_k - \bar{x})^2$$

with $|R_{x,y}| \leq |R_{1,\beta}|$ and $R_{1,\beta}$ and $R_{x,y}$ are, respectively, the correlation coefficients between x and $x^{\beta_{12}}$ and between x and y. The latter is unknown but often some decent guess about it is available.

The approximation of σ^2 in (5.6) is more convenient than the one in (4.2) as now we have that (5.4) is approximated by

$$\mathrm{MSE}_{\xi \mathrm{p}} \left(\hat{t}_{\mathrm{greg}} \right) \approx \beta_{11}^2 \left[\mathrm{MSE}_{\mathrm{p}} \left(\sum_s \frac{v_k}{\pi_k} \right) + F_0 \left(\sum_U \frac{g(x_k \mid \beta)^2}{\pi_k} - \sum_U g(x_k \mid \beta)^2 \right) \right] \tag{5.7}$$

with v_k given by (5.5). This expression depends neither on the intercept β_{01} nor the parameter σ, and the slope β_{11} becomes a proportionality constant that can be ignored.

The risk measure

As in Section 4, the asymptotic model expected MSE of the GREG estimator given by Result 2 can be seen as the loss incurred by assuming that δ is the true parameter when it is, in fact, β. Assuming a prior distribution on β, the risk (4.1) can be calculated.

6. Numerical examples

In Sections 2 and 3 we have established that the strategy $\pi\mathrm{ps}(\delta_2) - \mathrm{diff}(\delta_1)$ is optimal under a superpopulation model, but it is not robust to misspecifications of this model. In Subsection 6.1 we present a small Monte Carlo simulation study carried out to illustrate these results by comparing the optimal strategy and three alternatives.

In Sections 4 and 5 we introduced a measure that allows for quantifying the risk of implementing a sampling design, so allowing to guide the choice of design. In Subsection 6.2 we illustrate the use of the risk measure with real survey data.

6.1 Simulation study under a misspecified model

We compare the efficiency and robustness of four strategies through a simulation study. The strategies to be compared are $\pi\mathrm{ps}$ together with the difference estimator (which is optimal when the model is correct), $\pi\mathrm{ps}$ together with the GREG estimator (optimal design), stratified simple random sampling

(STSI) together with the difference estimator (optimal estimator) and STSI together with the GREG estimator.

Our implementation of πps makes use of Pareto πps (Rosén, 1997). There is a host of other schemes for drawing πps samples. Nevertheless, Pareto πps is a convenient method with good properties, see for example Rosén (2000).

Our implementation of STSI makes use of model-based stratification (Wright, 1983). We consider $H = 5$ strata with boundaries defined using Dalenius and Hodges (1959) cum$\sqrt{f}$-rule on $g(x_k \mid \delta_2)$ which is well described in (Särndal et al., 1992, page 463) and the sample is allocated using Neyman allocation, $n_h \propto N_h S_{gh}$. Using the cum$\sqrt{f}$-rule may be suboptimal (see Särndal et al., 1992, page 464) but the efficiency of stratification by a continuous size variable is fairly insensitive to the exact choice of boundaries.

We consider only misspecification of the spread. The trend term is of the form $f(x_k \mid \beta_1) = \beta_{10} + \beta_{11} x_k^{\beta_{12}}$ with $\beta_{10} = 1{,}000$, $\beta_{11} = 1$ and $\beta_{12} = 0.75$, 1 and 1.25. The true spread is $g(x_k \mid \beta_2) = x_k^{\beta_2}$ with $\beta_2 = 0.5$, 0.75 and 1. The working spread is $g(x_k \mid \delta_2) = x_k^{\delta_2}$ with $\delta_2 = 0.5$, 0.75 and 1.

We will use the difference estimator (2.1) calibrated on $f(x_k \mid \beta_1)$. Regarding the GREG estimator, we will fix β_{12}, whereas the coefficients β_{10} and β_{11} will be estimated.

The simulation is set out as follows. The population size is $N = 5{,}000$. The x-values are independent realizations from a gamma distribution with shape $\alpha = 4/100$ and scale $\lambda = 1{,}200$ plus one unit, whereas y_k is a realization from a gamma distribution with shape and scale

$$\alpha_k = \frac{\left(\beta_{10} + \beta_{11} x_k^{\beta_{12}}\right)^2}{\sigma_0^2 x_k^{2\beta_2}} \quad \text{and} \quad \lambda_k = \frac{\sigma_0^2 x_k^{2\beta_2}}{\beta_{10} + \beta_{11} x_k^{\beta_{12}}},$$

where σ^2 was set in such a way that the correlation between x and y is $\rho = 0.95$. The design MSE of a sample of size $n = 500$ is then computed for each strategy. Holding the x-values fixed, the process is iterated $B = 5{,}000$ times.

Table 6.1 shows the results of the simulation study. The first three columns indicate the model parameters. The fourth column shows the (simulated) model expected MSE of the strategy πps – dif, whereas the last three columns show the (simulated) efficiency of the strategies πps – GREG, STSI – dif and STSI – GREG compared to πps – dif (as a percentage), with efficiency defined as eff = $\mathrm{MSE}_{\xi,\pi\mathrm{ps}}(\hat{t}_y) / \mathrm{MSE}_{\xi,p}(\hat{t}_y)$ where the model expected MSEs are approximated by their simulated counterparts,

$$\mathrm{MSE}_{\xi,\mathrm{p}}(\hat{t}_y) = \mathrm{E}_\xi \mathrm{MSE}_\mathrm{p}(\hat{t}_y) \approx \frac{1}{B}\sum_{r=1}^{B} \mathrm{MSE}_\mathrm{p}^{(r)}(\hat{t}_y),$$

in such a way that a value of 100 indicates that the strategy is as efficient as πps – dif and values smaller (larger) than 100 indicate that the strategy is less (more) efficient than πps – dif.

The upper part of Table 6.1 shows the case when the working model coincides with the true model. As expected, the strategy that couples πps with the difference estimator (πps – dif) was always more efficient than the remaining strategies. Nevertheless, the loss in efficiency due to estimating some parameters through the GREG estimator is negligible. On the other hand, there is a remarkable loss in efficiency due to the use of STSI instead of πps. Finally, it is noted from (2.6) that as the anticipated MSE for all strategies does not depend on the trend f but only on the spread g, the efficiency remains constant under the same value of δ_2, independently of the value of β_{12}.

Table 6.1
Efficiency of three strategies as a percentage of the model expected MSE of πps – dif

Correct model						
β_{12}	β_2	δ_2	πps – dif	πps – GREG	STSI – dif	STSI – GREG
0.75	0.50	0.50	$2.78 \cdot 10^5$	99.9	57.3	57.3
0.75	0.75	0.75	$4.82 \cdot 10^4$	99.6	77.9	77.9
0.75	1.00	1.00	$1.90 \cdot 10^4$	99.0	83.2	83.2
1.00	0.50	0.50	$7.64 \cdot 10^6$	99.9	57.3	57.3
1.00	0.75	0.75	$7.20 \cdot 10^5$	99.7	77.9	77.9
1.00	1.00	1.00	$2.14 \cdot 10^5$	99.1	83.1	83.1
1.25	0.50	0.50	$1.46 \cdot 10^8$	99.9	57.3	57.3
1.25	0.75	0.75	$7.85 \cdot 10^6$	99.7	77.9	78.0
1.25	1.00	1.00	$1.81 \cdot 10^6$	99.2	83.1	83.1
Misspecified model						
β_{12}	β_2	δ_2	πps – dif	πps – GREG	STSI – dif	STSI – GREG
0.75	0.50	0.75	$3.98 \cdot 10^5$	99.9	98.9	98.9
0.75	0.75	1.00	$6.45 \cdot 10^4$	99.5	114.5	114.4
0.75	1.00	0.50	$4.73 \cdot 10^4$	100.1	133.9	134.0
1.00	0.50	1.00	$2.14 \cdot 10^7$	99.9	185.6	185.6
1.00	0.75	0.50	$1.03 \cdot 10^6$	100.1	93.1	93.2
1.00	1.00	0.75	$2.77 \cdot 10^5$	99.8	88.9	89.0
1.25	0.50	0.75	$2.09 \cdot 10^8$	99.9	98.9	98.9
1.25	0.75	1.00	$1.05 \cdot 10^7$	99.6	114.5	114.5
1.25	1.00	0.50	$4.50 \cdot 10^6$	100.3	134.0	134.2

The lower part of Table 6.1 shows some comparisons under a misspecified model, in particular, a misspecified spread. It can be noted that even under this mild misspecification of the model, πps – dif is not necessarily the best strategy anymore as the strategies using STSI were more efficient in several cases. However, it is not evident when will STSI be more efficient than πps or vice versa. The risk measure introduced in Section 4 can be used to guide the choice between designs. The results shown in this section agree with those shown by for example Holmberg and Swensson (2001).

6.2 Using the risk measure for choosing the design in a real survey

In this subsection we illustrate the implementation of the risk measure using data from a real survey. We want to estimate $t_y = \sum_U y_k$ where U is the set of residential properties in Bogotá, Colombia (of size N = 681,276) and y_k is the value of the k^{th} property in 2017 in COP. x_k, the built-up area of the

k^{th} property in square meters, is known for every $k \in U$. The auxiliary variable x has mean 184, standard deviation 110 and skewness 2.57. The desired sample size is n = 1,000.

We assume that a model of the type ξ_0 with $f(x_k \mid \delta_1) = \delta_{10} + \delta_{11} x_k^{\delta_{12}}$ and $g(x_k \mid \delta_2) = x_k^{\delta_2}$ adequately describes the association between x and y. We plan to use the GREG estimator for estimating δ_{10} and δ_{11}, i.e., $\delta_1^{**} = (\delta_{10}, \delta_{11})$. As this model has the form shown in Example 4, the model expected MSE can be approximated by expression (5.7).

We will use the risk (4.1) in order to assist the decision between πps or STSI using H = 6 strata. We take $h(\beta_{12}, \beta_2)$ as a bivariate normal distribution with no correlation between β_{12} and β_2. The integral is approximated using package `cubature` (Narasimhan, Johnson, Hahn, Bouvier and Kiêu, 2019) developed for the statistical software environment `R` (R Core Team, 2020).

We consider two cases with different degrees of confidence regarding the working model.

Case 1. In this case no information about δ_{12}, δ_2 or $R_{x,y}$ is available. Naive values of δ_{12} = 1, δ_2 = 1 and $R_{x,y}$ = 0.75 are considered. In order to reflect the uncertainty, $h(\beta)$ should have a large variance, therefore we set

$$\begin{bmatrix} \beta_{12} \\ \beta_2 \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} 1.0 \\ 1.0 \end{bmatrix}, \begin{bmatrix} 0.3295^2 & 0 \\ 0 & 0.3295^2 \end{bmatrix} \right).$$

The variance was chosen in such a way that 99% of the mass lies in the circle of radius 1. Evaluation of (4.1) yields $R(\pi\text{ps}) = 6.89 \cdot 10^{15} \beta_{11}^2$ and $R(\text{STSI}) = 1.59 \cdot 10^{15} \beta_{11}^2$, suggesting that a stratified design should be used.

The design MSE of both strategies is computed and we get, $\text{MSE}_{\pi\text{ps}}(\hat{t}_{\text{greg}}) = 2.29 \cdot 10^{25}$ and $\text{MSE}_{\text{STSI}}(\hat{t}_{\text{greg}}) = 1.36 \cdot 10^{25}$. The strategy suggested by (4.1) was indeed the best choice.

Case 2. Using a sample from 2010, prior values of δ_{12} = 1.9, δ_2 = 2 and $R_{x,y}$ = 0.7 are proposed. As the uncertainty here is smaller than that in Case 1, we set a smaller variance,

$$\begin{bmatrix} \beta_{12} \\ \beta_2 \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} 1.9 \\ 2.0 \end{bmatrix}, \begin{bmatrix} 0.2471^2 & 0 \\ 0 & 0.2471^2 \end{bmatrix} \right).$$

The variance was chosen in such a way that 99% of the mass lies in the circle of radius 0.75. Evaluation of (4.1) yields $R(\pi\text{ps}) = 7.08 \cdot 10^{22} \beta_{11}^2$ and $R(\text{STSI}) = 4.06 \cdot 10^{18} \beta_{11}^2$, suggesting that a stratified design should be used.

The design MSE of both strategies is computed and we get $\text{MSE}_{\pi\text{ps}}(\hat{t}_{\text{greg}}) = 1.85 \cdot 10^{28}$ and $\text{MSE}_{\text{STSI}}(\hat{t}_{\text{greg}}) = 1.91 \cdot 10^{25}$. Note that the use of (4.1) prevented us from using πps, whose MSE is almost one thousand times bigger than the one under stratified sampling!

7. Conclusions

The strategy that couples πps with the difference estimator is optimal when the parameters of the superpopulation model are known. Taking into account that these assumptions are seldom satisfied, it was shown in Section 3 and illustrated in Subsection 6.1 that this optimality breaks down even under small misspecifications of the model.

In Section 4 we propose a method for choosing the sampling design, which is extended to its use with the GREG estimator in Section 5. The method allows for taking the uncertainty about the model parameters into account by introducing a prior distribution on them. Although it could be argued that a source of subjectivity is added by introducing a prior distribution on the parameters, our view is that it is more subjective to choose the design without any type of assessment of the assumptions. Furthermore, inference is still design-based, as the prior is used only for choosing the design.

The method was illustrated with a real dataset, yielding satisfactory results. It should be noted that although the illustrations used stratified simple random sampling, the method in this article is valid for any sampling design.

Appendix

Proof of (4.2)

Proof. The following expectations are required in the proof,

$$\mathrm{E}_\xi Y_k = \mathrm{E}_\xi [f(x_k|\beta_1) + \varepsilon_k] = f(x_k|\beta_1) \tag{A.1}$$

$$\mathrm{E}_\xi Y_k^2 = \mathrm{E}_\xi \left[(f(x_k|\beta_1) + \varepsilon_k)^2 \right] = f(x_k|\beta_1)^2 + \sigma^2 g(x_k|\beta_2)^2 \tag{A.2}$$

$\mathrm{E}_\xi \bar{Y}$, $\mathrm{E}_\xi \bar{Y}^2$ and $\mathrm{E}_\xi \overline{fY}$ are obtained using (A.1) and (A.2),

$$\mathrm{E}_\xi \bar{Y} = \mathrm{E}_\xi \left[\frac{1}{N} \sum_U Y_k \right] = \frac{1}{N} \sum_U \mathrm{E}_\xi Y_k = \frac{1}{N} \sum_U f(x_k|\beta_1) \equiv \bar{f} \tag{A.3}$$

$$\mathrm{E}_\xi \bar{Y}^2 = \mathrm{E}_\xi \left[\frac{1}{N} \sum_U Y_k^2 \right] = \frac{1}{N} \sum_U \left(f(x_k|\beta_1)^2 + \sigma^2 g(x_k|\beta_2)^2 \right) \equiv \bar{f}^2 + \sigma^2 \bar{g}^2 \tag{A.4}$$

$$\mathrm{E}_\xi \overline{fY} = \mathrm{E}_\xi \left[\frac{1}{N} \sum_U f(x_k|\beta) Y_k \right] = \frac{1}{N} \sum_U f(x_k|\beta) \mathrm{E}_\xi Y_k = \frac{1}{N} \sum_U f(x_k|\beta)^2 = \bar{f}^2. \tag{A.5}$$

Now, using (A.3), (A.4) and (A.5) we get

$$\mathrm{E}_\xi \left[\overline{fY} - \overline{fY} \right] = \bar{f}^2 - \bar{f}^2 = S_{f,f} \tag{A.6}$$

$$\mathrm{E}_\xi \left[\bar{Y}^2 - \bar{Y}^2 \right] = \bar{f}^2 + \sigma^2 \bar{g}^2 - \bar{f}^2 = S_{f,f} + \sigma^2 \bar{g}^2. \tag{A.7}$$

Using (A.6) and (A.7), we obtain an approximation to the correlation coefficient, $R_{f,y}$,

$$R_{f,y}^2 = \frac{\left(\overline{fy} - \overline{f}\,\overline{y}\right)^2}{\left(\overline{f}^2 - \overline{f}^2\right)\left(\overline{y}^2 - \overline{y}^2\right)} \approx \frac{\mathrm{E}_\xi^2\left[\overline{fY} - \overline{f}\,\overline{Y}\right]}{\mathrm{E}_\xi\left[\left(\overline{f}^2 - \overline{f}^2\right)\left(\overline{Y}^2 - \overline{Y}^2\right)\right]} = \frac{S_{f,f}}{S_{f,f} + \sigma^2 \overline{g}^2}. \qquad \text{(A.8)}$$

Solving (A.8) for σ^2 we get (4.2), as desired. The proof of (5.6) is analogous.

References

Beaumont, J.-F., Haziza, D. and Ruiz-Gazen, A. (2013). A unified approach to robust estimation in finite population sampling. *Biometrika*, 100, 3, 555-569.

Bramati, M. (2012). Robust Lavallée-Hidiroglou stratified sampling strategy. *Survey Research Methods*, 6, 3, 137-143.

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

Cassel, C.M., Särndal, C.-E. and Wretman, J. (1977). *Foundations of Inference in Survey Sampling*. New York: John Wiley & Sons, Inc.

Dalenius, T., and Hodges, J.L. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, 54, 88-101.

Godambe, V.P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society, Series B*, 17, 269-278.

Hájek, J. (1959). Optimal strategy and other problems in probability sampling. *Casopis Pro Pestování Matematiky*, 84, 4, 387-423.

Holmberg, A., and Swensson, B. (2001). On pareto πps sampling: Reflections on unequal probability sampling strategies. *Theory of Stochastic Processes*, 7(23), 142-155.

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

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

Lanke, J. (1973). On UMV-estimators in survey sampling. *Metrika*, 20, 196-202.

Narasimhan, B., Johnson, S., Hahn, T., Bouvier, A. and Kiêu, K. (2019). *Cubature: Adaptive Multivariate Integration Over Hypercubes*. R package version 2.0.4. https://CRAN.R-project.org/package=cubature.

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

R Core Team (2020). R: A language and environment for statistical computing. *The R Foundation for Statistical Computing*, Vienna, Austria. URL https://www.R-project.org/.

Rosén, B. (1997). On sampling with probability proportional to size. *Journal of Statistical Planning and Inference*, 62, 159-191.

Rosén, B. (2000). *Generalized Regression Estimation and Pareto πps*. R&D Report 2000:5. Statistics Sweden.

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

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

Tillé, Y., and Wilhelm, M. (2017). Probability sampling designs: Principles for choice of design and balancing. *Statistical Science*, 32(2), 176-189.

Wright, R.L. (1983). Finite population sampling with multivariate auxiliary information. *Journal of the American Statistical Association*, 78, 879-884.

Zhai, Z., and Wiens, D. (2015). Robust model-based stratification sampling designs. *The Canadian Journal of Statistics*, 43, 4, 554-577.

Survey Methodology, June 2021
Vol. 47, No. 1, pp. 91-122
Statistics Canada, Catalogue No. 12-001-X

Bayesian predictive inference of small area proportions under selection bias

Seongmi Choi, Balgobin Nandram and Dalho Kim[1]

Abstract

In a previous paper, we developed a model to make inference about small area proportions under selection bias in which the binary responses and the selection probabilities are correlated. This is the homogeneous nonignorable selection model; nonignorable selection means that the selection probabilities and the binary responses are correlated. The homogeneous nonignorable selection model was shown to perform better than a baseline ignorable selection model. However, one limitation of the homogeneous nonignorable selection model is that the distributions of the selection probabilities are assumed to be identical across areas. Therefore, we introduce a more general model, the heterogeneous nonignorable selection model, in which the selection probabilities are not identically distributed over areas. We used Markov chain Monte Carlo methods to fit the three models. We illustrate our methodology and compare our models using an example on severe activity limitation of the U.S. National Health Interview Survey. We also perform a simulation study to demonstrate that our heterogeneous nonignorable selection model is needed when there is moderate to strong selection bias.

Key Words: Biserial correlation; Metropolis-Hastings algorithm; Nonignorable selection model; Official statistics; Selection probabilities.

1. Introduction

In many complex sample surveys, individuals are sampled with differential selection probabilities. For binary responses, if the proportion of positive responses among the sampled individuals differs substantially from those among the nonsampled individuals, there is a selection bias. In some cases a selection bias is obtained by design (e.g., probability proportional to size (PPS) sampling) and we can take care of the selection bias, but in other problems, this is not the case. For example, in non-probability samples the selection probabilities are unknown. In our problem, we assume that the survey weights or selection probabilities can help us to understand the selection bias. We want to make inference about the finite population proportions of the small areas when a possibly biased sample is available from each area. Choi, Nandram and Kim (2017), henceforth CNK, extended the model of Nandram, Bhatta, Bhadra and Shen (2013), henceforth NBBS, who studied a single area, to accommodate inference about small areas. While the CNK model assumes that distribution of the selection probabilities is the same across areas, our new contribution is to assume that the distributions of the selection probabilities over areas are different, but they share an effect.

There are two types of models that can be considered when making inference about small areas. First, we can use an ignorable selection model in which the response variable is not related to the selection probabilities. Such a model will not adjust for the selection bias, and will produce biased estimates if there

1. Seongmi Choi, Health Insurance Policy Research Istitute, National Health Insurance Service, Wonju, Gangwon 26464, South Korea. E-mail: edisil1202@gmail.com; Balgobin Nandram, Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609, USA. E-mail: balnan@wpi.edu; Dalho Kim, Department of Statistics, Kyungpook National University, Daegu 41566, South Korea. E-mail: dalkim@knu.ac.kr.

are no other covariates to act like the selection probabilities. We assume that the only available information is the sampled responses, sampled selection probabilities and area identifiers. Second, in a nonignorable selection model, the response variables are related to the selection probabilities. In our study on binary responses, the distribution of the selection probabilities for the "yes" responses is different from that for the "no" responses, thereby making the response variables and selection probabilities correlated. Official statistics are obtained from many complex surveys, and these surveys are conveniently designed with selection bias (PPS sampling is usually a part of a complex survey design). This study is important in the construction of official statistics from complex surveys.

It is pertinent to give a brief discussion on recent developments. For continuous responses, Pfeffermann (1988), Pfeffermann (1993), Sverchkov and Pfeffermann (2004) and Pfeffermann, Krieger and Rinott (1998) specified a relation between the survey weights and the response variable. Chambers, Dorfman and Wang (1998) assumed that the selection probabilities are related to the continuous responses. A related study within the Bayesian paradigm is given by Ma, Sedransk, Nandram and Chen (2018). There is also an application of this method to calculate the total gas consumption in the US using a PPS sample; see Nandram, Choi, Shen and Burgos (2006). Chen, Elliott and Little (2010) used penalized spline to make a Bayesian predictive inference for PPS sampling. However, because these approaches require some information about the nonsampled selection probabilities, they are difficult to use. The nonsampled selection probabilities are not available to secondary data analysts and we continue to assume that these nonsampled selection probabilities are not available in our work. Pfeffermann and Sverchkov (2007) extended the work of Sverchkov and Pfeffermann (2004) to accommodate small areas with informative sampling of areas and within selected areas. Like Chen et al. (2010), Opsomer, Glaeskens, Ranalli, Kauermann and Breidt (2008) used a penalized spline regression to construct a small-area model, which includes the selection probabilities, in a non-parametric manner. While all these works are for continuous response in small areas, we analyze binary data in this paper.

Our approach to selection bias problems has been to adjust the sample part of a population model. That is, a model is constructed for the entire population, the model is then adjusted to accommodate the sample with the selection probabilities, and the superpopulation parameters are then estimated. Once this is done for the sample, prediction is done for the entire finite population. This approach to selection bias was described nicely by Malec, Davis and Cao (1999). NBBS reviewed and provided a full Bayesian analyis to the method of Malec et al. (1999); Nandram (2007) provided a surrogate sampling procedure within the spirit of Malec et al. (1999). There are many extensions to this approach to accommodate small areas. For example, to analyze continuous responses, Nandram and Choi (2010) included the selection probabilities into a full Bayesian nonignorable model in order to analyze continuous body mass index data, which were discretized in an elegant manner. Malec et al. (1999) used a hierarchical Bayesian model to accommodate a selection mechanism for binary data. As stated by NBBS, there are two possible problems with this model, "First, the θ_{uy} are only weakly identified. Second, the parameters θ_{uy} are never known, and in a Bayesian framework these must also be stochastic. In this paper, in a single attempt, we show how to

solve these two problems (weak identifiability and stochastic parameters) for a biased sample drawn from a binary population using information from the survey weights (or selection probabilities)". Using a nonignorable selection model, NBBS showed how to correct for these two problems for an individual-area model.

Recently, there has been some interesting Bayesian activities for PPS sampling with continuous responses. Zangeneh and Little (2015) used a Bayesian bootstrap to model the measures of size (related to the selection probabilities) and a spline regression of the responses conditional on the measures of size. For the same problem, Si, Pillai and Gelman (2015) poststratified the selection probabilities and used a Gaussian process to model the responses. These are good approaches to Bayesian analyzes of PPS sampling and can be applied to our problem. However, our approach is different because we do not model the nonsampled selection probabilities, rather we adjust a population model.

CNK studied a special case for an extension of NBBS model within the small area context. They assumed that the sample selection probabilities have the same support over the set π_u^* over areas, and the distribution of the selection probabilities given the binary response y_{ij} is

$$P\left(\pi_{ij} = \pi_u^* \,\middle|\, \boldsymbol{\theta},\, y_{ij} = y\right) = \theta_{uy},\, u = 1, \ldots, U,\, y = 0, 1,\, j = 1, \ldots, n_i,\, i = 1, \ldots, \ell.$$

This model assumes that the sample selection probabilities have the same support and each area has the same distribution for the selection probabilities; homogeneous nonignorable selection model. However, the homogeneity assumption is strong and might not be true for most of real settings. In practice, the distribution of the sample selection probabilities can be very different across domains due to sampling designs. The distribution of the sample selection probabilities given $\mathbf{y}$ can also vary a lot across domains. In this paper, we consider a model under a heterogeneity assumption, which is that the sample selection probabilities have different supports and different distributions over areas; this is the heterogeneous nonignorable selection model.

CNK used a simulation study to show how different a baseline ignorable selection model and the homogeneous nonignorable selection model can be when there is selection bias. But it is well known that design information needs to be included when the sample is not randomly selected. In this paper, we use a simulation study to assess the performance of our heterogeneous nonignorable selection model. We draw data from the homogeneous nonignorable selection model and heterogeneous nonignorable selection model respectively, and fit the three models. Then we compare the performance of the models using several measures.

In this paper, we consider the problem of making inference about the finite population proportions of the small areas when there is likely to be a different selection bias by areas. Specifically we extend the homogeneous nonignorable selection model of CNK to accommodate selection probabilities that have different supports in different areas. In Section 2, we give a review of the ignorable selection model and the homogeneous nonignorable selection model, which were studied previously. In Section 3, we show

how to adjust a Bayesian ignorable selection model to incorporate the selection bias in a small area framework when the sample selection probabilities have different distributions by areas. We also describe how to perform the computation. In Section 4, we provide results of an illustrative example on severe activity limitation in the 1995 National Health Interview Survey. We also provide a simulation study to assess the performance of the heterogeneous nonignorable selection model. Section 5 has summary and concluding remarks.

2. Review

This section gives a review of the ignorable selection (IS) model discussed by NBBS and the homogeneous nonignorable selection (HoS) model discussed by CNK. We assume that there are ℓ areas and the population size of the i^{th} area is N_i, $i = 1, \ldots, \ell$, which are known. We consider binary responses.

The samples taken from each area can be biased in that the proportion of positive responses among the sampled units may be different from the proportion of positive responses among the nonsampled units. A sample of $n_i << N_i$ is taken from the i^{th} area, and j^{th} unit within the i^{th} area is taken with selection probabilities π_{ij}, $j = 1, \ldots, N_i$, $i = 1, \ldots, \ell$. As common to many problems of this kind, the selection probabilities are observed only for the sampled values. (Only the sampled selection probabilities are presented to secondary data users.) Design scientists adjust the selection probabilities to take care of nonresponse and other nonsampling errors. But the selection probabilities are a major part of the survey weights and we assume that the selection probabilities are approximately the reciprocal of the survey weights in our application.

Let $y_{i1}, \ldots, y_{iN_i}$, $i = 1, \ldots, \ell$, denote the binary responses in the ℓ areas. A biased sample S_i of size n_i is available from the i^{th} area together with the selection probabilities of the sample. Denote the sampled values by $y_{i1}, \ldots, y_{in_i}$, $i = 1, \ldots, \ell$, and the set of nonsampled values by $\bar{S}_i$. Inference is required for the finite population proportion for each area. Let $P_i = \sum_{j=1}^{N_i} y_{ij} / N_i$ denote the small area proportion of the i^{th} area and $\hat{p}_i = \sum_{j=1}^{n_i} y_{ij} / n_i$ denote the corresponding sample proportions. Clearly, $\hat{p}_i$ can be a biased estimator of P_i, the proportion of ones in the finite population. In design-based survey analysis, P_i are fixed unknown quantities, but in the Bayesian paradigm, P_i are random variables, which are to be predicted (i.e., Bayesian predictive inference).

We review the IS model in Section 2.1 and the HoS model in Section 2.2.

2.1 Ignorable selection model

A standard ignorable selection model for the binary variables y_{ij}, $j = 1, \ldots, N_i$, $i = 1, \ldots, \ell$, is

$$y_{ij} \mid p_i \overset{\text{ind}}{\sim} \text{Bernoulli}(p_i),$$

$$p_i \mid \mu, \tau \overset{\text{iid}}{\sim} \text{Beta}(\mu\tau, (1-\mu)\tau)$$

and

$$p(\mu, \tau) = \frac{1}{(1+\tau)^2}, \tau \geq 0 (0 < \mu < 1)$$

(i.e., a priori μ and τ are independent). Here the prior for τ is proper and noninformative.

For notational convenience, let $\mathbf{p} = (p_1, \ldots, p_\ell)$, $\mathbf{y} = \left(y_{11}, \ldots, y_{1n_1}, \ldots, y_{\ell 1}, \ldots, y_{\ell n_\ell}\right)$ and $s_i = \sum_{j=1}^{n_i} y_{ij}$. Then Bayesian predictive inference of P_i can be performed based on the following posterior distribution,

$$p(y_{ij} \mid \mathbf{y}) = \int p(y_{ij} \mid p_i, \mathbf{y}) \pi(p_i \mid \mathbf{y}) dp_i = \int p(y_{ij} \mid p_i) \pi(p_i \mid \mathbf{y}) dp_i, \; j = n_i + 1, \ldots, N_i, \; i = 1, \ldots, \ell,$$

where $y_{ij} \mid p$ are independent, $\pi(p_i \mid \mathbf{y}) = \int \pi(p_i \mid \mu, \tau, \mathbf{y}) \pi(\mu, \tau \mid \mathbf{y}) d\mu d\tau$ and $p_i \mid \mu, \tau, \mathbf{y}$ are independent with $p_i \mid \mu, \tau, \mathbf{y} \sim \text{Beta}(s_i + \mu\tau, n_i - s_i + (1-\mu)\tau)$.

2.2 Homogeneous nonignorable selection model

CNK studied a special case for an extension of NBBS model within the small area context. They assumed that the sample selection probabilities $(\pi_{i1}, \ldots, \pi_{in_i})$ have the same support over the set $\pi_u^*, u = 1, \ldots, U$ for $i = 1, \ldots, \ell$.

Letting $\boldsymbol{\theta} = (\theta_{10}, \ldots, \theta_{U0}, \theta_{11}, \ldots, \theta_{U1}) = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)$, (say) and $\boldsymbol{\pi} = \left(\pi_{11}, \ldots, \pi_{1n_1}, \ldots, \pi_{\ell 1}, \ldots, \pi_{\ell n_\ell}\right)$, the distribution of the selection probabilities, given the binary response y_{ij}, is

$$P(\pi_{ij} = \pi_u^* \mid \boldsymbol{\theta}, y_{ij} = y) = \theta_{uy}, u = 1, \ldots, U, y = 0, 1, j = 1, \ldots, n_i, i = 1, \ldots, \ell$$

and

$$y_{ij} \mid p_i \overset{\text{iid}}{\sim} \text{Bernoulli}(p_i), \; j = 1, \ldots, n_i, \; i = 1, \ldots, \ell.$$

It is worth noting here that the θ_{uy} are not selection probabilities.

Note that the sample selection probabilities have the same support and distribution by areas. However, this assumption is strong and might not be true for most of real settings. In practice, the distribution of the sample selection probabilities can be very different across domains due to sampling designs.

To proceed, they make a one-to-one transformation from p_i to q_i via

$$q_i = \frac{a_1 p_i}{a_1 p_i + a_0 (1 - p_i)}, \; i = 1, \ldots, \ell.$$

Then, they assumed that $\mathbf{q}$, $\boldsymbol{\theta}_0$, $\boldsymbol{\theta}_1$ are independent with

$$q_i \mid \mu, \tau_1 \overset{\text{iid}}{\sim} \text{Beta}(\mu\tau_1, (1-\mu)\tau_1), \; i = 1, \ldots, \ell,$$

$$\boldsymbol{\theta}_0 \,|\, \tau_0 \sim \text{Dirichlet}\left(\boldsymbol{\theta}_0^{(0)}\tau_0\right) \quad \text{and} \quad \boldsymbol{\theta}_1 \,|\, \tau_0 \sim \text{Dirichlet}\left(\boldsymbol{\theta}_1^{(0)}\tau_0\right),$$

where $\boldsymbol{\theta}_0^{(0)}$ and $\boldsymbol{\theta}_1^{(0)}$ are to be specified.

A priori they assumed

$$p(\mu, \tau_0, \tau_1) = \frac{1}{(1+\tau_0)^2}\,\frac{1}{(1+\tau_1)^2},\ 0 < \mu < 1,\ \tau_0, \tau_1 \geq 0.$$

Once they drew a sample from the posterior $\pi(\mathbf{q}, \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \mu, \tau_0, \tau_1 \,|\, \mathbf{y})$, by retransforming from q_i to p_i,

$$p_i = \frac{a_0 q_i}{a_0 q + a_1 (1 - q_i)},$$

they obtained a sample from the posterior distribution of $\pi(\mathbf{p}, \boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \mu, \tau_0, \tau_1 \,|\, \mathbf{y})$.

Finally, one can draw the entire finite population values, y_{ij}, from $\text{Bernoulli}(p_i)$, $i = 1, \ldots, \ell$, $j = 1, \ldots, n_i$ independently, using estimated $\mathbf{p}$, and can make an inference about the small area proportions. The distribution of the sample selection probabilities given $\mathbf{y}$ can also vary a lot across domains. Therefore, in Section 3, we consider a model which has different supports and distributions for the sample selection probabilities by areas.

3. Heterogeneous nonignorable selection model

In Section 3.1, we describe the heterogeneous nonignorable selection (HeS) model. We also show how to perform the computations in Section 3.2. We show how to fit this model and how to make inference about the small area proportions. Inference about the small area proportions under the HeS model is obtained using surrogate samples (Nandram, 2007).

3.1 Methodology

We assume that the sample selection probabilities $(\pi_{i1}, \ldots, \pi_{in_i})$ have the different supports over the set $\pi_{iu}^*, u = 1, \ldots, U_i$, for $i = 1, \ldots, \ell$. That is, π_{ij}, $j = 1, \ldots, n_i$ have a histogram where the midpoints of the categories are the π_{iu}^*, for $i = 1, \ldots, \ell$. These π_{iu}^* are assumed known and the π_{ij} are assumed to be random quantities. For notational convenience, let $\boldsymbol{\theta}_i = (\theta_{i10}, \ldots, \theta_{iU_i 0}, \theta_{i11}, \ldots, \theta_{iU_i 1}) = (\boldsymbol{\theta}_{i0}, \boldsymbol{\theta}_{i1})$ (say), $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_\ell)$, and $\boldsymbol{\pi} = (\pi_{11}, \ldots, \pi_{1n_1}, \ldots, \pi_{\ell 1}, \ldots, \pi_{\ell n_\ell})$. The distribution of the selection probabilities, given the binary response y_{ij}, is

$$P(\pi_{ij} = \pi_{iu}^* \,|\, \boldsymbol{\theta}, y_{ij} = y) = \theta_{iuy},\ u = 1, \ldots, U_i,\ y = 0, 1,\ j = 1, \ldots, n_i,\ i = 1, \ldots, \ell$$

and

$$y_{ij} \,|\, p_i \overset{\text{iid}}{\sim} \text{Bernoulli}(p_i),\ j = 1, \ldots, N_i,\ i = 1, \ldots, \ell.$$

Again, it is worth noting that the θ_{iuy} are not selection probabilities.

To accommodate the sample selection scheme, we assume that θ_{iu0} and θ_{iu1}, for $i = 1, \ldots, \ell$, are different. Note that we consider the heterogeneity assumption for the sample selection probabilities. We replace the homogeneity assumption with the heterogeneity assumption for the sample selection probabilities of the HoS model, so that the sample selection probabilities have different supports and the distributions of the selection probabilities are different by areas.

Let δ_{ij}, π_{ij} and y_{ij} denote the selection indicator, the selection probability and the binary response of the j^{th} unit in the i^{th} small area in the population respectively. Essentially, NBBS postulated that the $(\delta_{ij}, \pi_{ij}, y_{ij})$ within the i^{th} small area are independently distributed with

$$\begin{aligned} P(\delta_{ij} = \delta, \pi_{ij} = \pi_{iu}^*, y_{ij} = y \,|\, \boldsymbol{\theta}_i, p_i) &= P_1(\delta_{ij} = \delta \,|\, \pi_{ij} = \pi_{iu}^*) P_2(\pi_{ij} = \pi_{iu}^* \,|\, y_{ij} = y, \boldsymbol{\theta}_i) P_3(y_{ij} = y \,|\, p_i) \\ &= (\pi_{iu}^*)^{\delta} (1 - \pi_{iu}^*)^{1-\delta} \theta_{iuy} p_i^y (1 - p_i)^{1-y}, \\ &\quad \delta = 0, 1, u = 1, \ldots, U_i, y = 0, 1, j = 1, \ldots, n_i, i = 1, \ldots, \ell. \end{aligned}$$

Thus, there is a joint probability mass function for the selection indicator and the response indicator. Therefore, the model that NBBS specified is a nonignorable selection model (i.e., NBBS assumed that the selection mechanism is selection not at random (SNAR)). Since there are no data when $\delta = 0$ (i.e., π and y are both unobserved), NBBS used the conditional probability mass function

$$P(\pi_{ij} = \pi_{iu}^*, y_{ij} = y \,|\, \delta_{ij} = 1, \boldsymbol{\theta}_i, p_i) = \frac{\pi_{iu}^* \theta_{iuy} p_i^y (1 - p_i)^{1-y}}{\sum_{y=0}^{1} \sum_{u=1}^{U_i} \pi_{iu}^* \theta_{iuy} p_i^y (1 - p_i)^{1-y}}, \; i = 1, \ldots, \ell;$$

see the probability mass function in (4) of NBBS.

We have the data (π_{ij}, y_{ij}), $j = 1, \ldots, n_i$, $i = 1, \ldots, \ell$. Since the sampling units are independent, the likelihood function is given by

$$\prod_{i=1}^{\ell} \prod_{j=1}^{n_i} P(\pi_{ij} = \pi_{iu}^*, Y_{ij} = y_{ij} \,|\, \delta_{ij} = 1, \boldsymbol{\theta}_i, \mathbf{p}) = \prod_{i=1}^{\ell} \prod_{j=1}^{n_i} \frac{\pi_{iu}^* \theta_{iuy_{ij}} p_i^{y_{ij}} (1 - p_i)^{1-y_{ij}}}{\sum_{y_{ij}=0}^{1} \sum_{u=1}^{U_i} \pi_{iu}^* \theta_{iuy_{ij}} p_i^{y_{ij}} (1 - p_i)^{1-y_{ij}}},$$

where π_{iu}^*, $u = 1, \ldots, U_i$, $i = 1, \ldots, \ell$ are known. The likelihood function can be rewritten as

$$\begin{aligned} P(\mathbf{y}, \boldsymbol{\pi} \,|\, \boldsymbol{\theta}, \mathbf{p}) &= \frac{\prod_{i=1}^{\ell} \prod_{j=1}^{n_i} \pi_{iu}^* \theta_{iuy_{ij}} \prod_{i=1}^{\ell} \prod_{j=1}^{n_i} p_i^{y_{ij}} (1 - p_i)^{1-y_{ij}}}{\prod_{i=1}^{\ell} \prod_{j=1}^{n_i} \sum_{y_{ij}=0}^{1} \sum_{u=1}^{U_i} \pi_{iu}^* \theta_{iuy_{ij}} p_i^{y_{ij}} (1 - p_i)^{1-y_{ij}}} \\ &= \frac{\prod_{i=1}^{\ell} \prod_{u=1}^{U_i} (\pi_{iu}^* \theta_{iu0})^{g_{iu0}} \prod_{i=1}^{\ell} \prod_{u=1}^{U_i} (\pi_{iu}^* \theta_{iu1})^{g_{iu1}} \prod_{i=1}^{\ell} p_i^{s_i} (1 - p_i)^{n_i - s_i}}{\prod_{i=1}^{\ell} \left[p_i \sum_{u=1}^{U_i} \pi_{iu}^* \theta_{iu1} + (1 - p_i) \sum_{u=1}^{U_i} \pi_{iu}^* \theta_{iu0} \right]^{n_i}}, \end{aligned}$$

where $s_i = \sum_{j=1}^{n_i} y_{ij}$, g_{iu0} is the cell count for category u at $y = 0$ and g_{iu1} is the cell count for category u at $y = 1$ under the area i. Note that $\sum_{u=1}^{U_i} g_{iu0} = n_i - s_i$, $\sum_{u=1}^{U_i} g_{iu1} = s_i$ and $\sum_{u=1}^{U_i} (g_{iu0} + g_{iu1}) = n_i$. This likelihood includes the selection bias.

Let $a_{iy} = \sum_{u=1}^{U_i} \pi_{iu}^* \theta_{iuy}$, $y = 0, 1$, $i = 1, \ldots, \ell$. Differences between a_{i0} and a_{i1} for some i indicate that there is selection bias. The likelihood function can be expressed as

$$P(\mathbf{y}, \boldsymbol{\pi} | \boldsymbol{\theta}, \mathbf{p}) = \prod_{i=1}^{\ell} \prod_{u=1}^{U_i} (\pi_{iu}^* \theta_{iu0})^{g_{iu0}} \prod_{i=1}^{\ell} \prod_{u=1}^{U_i} (\pi_{iu}^* \theta_{iu1})^{g_{iu1}} \prod_{i=1}^{\ell} \frac{p_i^{s_i} (1 - p_i)^{n_i - s_i}}{[a_{i1} p_i + a_{i0} (1 - p_i)]^{n_i}}.$$

We make a one-to-one transformation from p_i to q_i via

$$q_i = \frac{a_{i1} p_i}{a_{i1} p_i + a_{i0} (1 - p_i)}.$$

Note that if θ_{iuy} does not depend on y, q_i and p_i are the same. In this case the likelihood would be the same as the ignorable case. Let $\mathbf{q} = (q_1, \ldots, q_\ell)$. Then the likelihood function can be expressed as

$$P(\mathbf{y}, \boldsymbol{\pi} | \boldsymbol{\theta}, \mathbf{q}) = \prod_{i=1}^{\ell} \frac{\prod_{u=1}^{U_i} (\pi_{iu}^* \theta_{iu0})^{g_{iu0}}}{\left(\sum_{u=1}^{U_i} \pi_{iu}^* \theta_{iu0}\right)^{n_i - s_i}} \prod_{i=1}^{\ell} \frac{\prod_{u=1}^{U_i} (\pi_{iu}^* \theta_{iu1})^{g_{iu1}}}{\left(\sum_{u=1}^{U_i} \pi_{iu}^* \theta_{iu1}\right)^{s_i}} \prod_{i=1}^{\ell} q_i^{s_i} (1 - q_i)^{n_i - s_i}.$$

We assume that $\mathbf{q}$, $\boldsymbol{\theta}_{i0}$, $\boldsymbol{\theta}_{i1}$ are independent, and we take

$$q_i | \mu, \tau_1 \overset{\text{iid}}{\sim} \text{Beta}(\mu\tau_1, (1 - \mu)\tau_1), \; i = 1, \ldots, \ell,$$

$$\boldsymbol{\theta}_{i0} | \tau_0 \overset{\text{ind}}{\sim} \text{Dirichlet}(\boldsymbol{\theta}_{i0}^{(0)} \tau_0) \quad \text{and} \quad \boldsymbol{\theta}_{i1} | \tau_0 \overset{\text{ind}}{\sim} \text{Dirichlet}(\boldsymbol{\theta}_{i1}^{(0)} \tau_0),$$

where $\boldsymbol{\theta}_{i0}^{(0)}$ and $\boldsymbol{\theta}_{i1}^{(0)}$ are to be specified. Recall that $\mathbf{x} | \boldsymbol{\mu}, \tau \sim \text{Dirichlet}(\boldsymbol{\mu}\tau)$ has the density $f(\mathbf{x} | \boldsymbol{\mu}, \tau) = \frac{\prod_{i=1}^{k} x_i^{\mu_i \tau - 1}}{D(\boldsymbol{\mu}\tau)}$, $0 < x_i < 1$, $\sum_{i=1}^{k} x_i = 1$, where $D(\boldsymbol{\mu}\tau) = \prod_{i=1}^{k} \Gamma(\mu_i \tau) / \Gamma(\tau)$, $0 < \mu_i < 1$, $\sum_{i=1}^{k} \mu_i = 1$, and $\tau > 0$.

Finally, a priori we assume

$$p(\mu, \tau_0, \tau_1) = \frac{1}{(1 + \tau_0)^2} \frac{1}{(1 + \tau_1)^2}, \; 0 < \mu < 1, \; \tau_0, \tau_1 \geq 0.$$

Of course one can use a half Cauchy density but these are very similar. This latter prior is used to avoid the difficulties associated with improper priors of the form $p(\tau) \propto 1/\tau$ (e.g., see Gelman, 2006).

Hence, the joint prior density of $\mathbf{q}$, $\boldsymbol{\theta}_{10}, \ldots, \boldsymbol{\theta}_{\ell 0}$, $\boldsymbol{\theta}_{11}, \ldots, \boldsymbol{\theta}_{\ell 1}$, μ, τ_0, τ_1 is

$$\pi(\mathbf{q}, \boldsymbol{\theta}_{10}, \ldots, \boldsymbol{\theta}_{\ell 0}, \boldsymbol{\theta}_{11}, \ldots, \boldsymbol{\theta}_{\ell 1}, \mu, \tau_0, \tau_1) \propto \prod_{i=1}^{\ell} \frac{\prod_{u=1}^{U_i} \theta_{iu0}^{\theta_{iu0}^{(0)} \tau_0 - 1}}{D(\boldsymbol{\theta}_{i0}^{(0)} \tau_0)} \prod_{i=1}^{\ell} \frac{\prod_{u=1}^{U_i} \theta_{iu1}^{\theta_{iu1}^{(0)} \tau_0 - 1}}{D(\boldsymbol{\theta}_{i1}^{(0)} \tau_0)}$$
$$\times \frac{\prod_{i=1}^{\ell} q_i^{\mu\tau_1 - 1} (1 - q_i)^{(1-\mu)\tau_1 - 1}}{[B(\mu\tau_1, (1 - \mu)\tau_1)]^{\ell}} \frac{1}{(1 + \tau_0)^2} \frac{1}{(1 + \tau_1)^2}.$$

Using Bayes' theorem, the joint posterior density of $\mathbf{q}$, $\boldsymbol{\theta}_{10}, \ldots, \boldsymbol{\theta}_{\ell 0}$, $\boldsymbol{\theta}_{11}, \ldots, \boldsymbol{\theta}_{\ell 1}$, μ, τ_0, τ_1, given the data $\mathbf{y}$, is

$$\pi(\mathbf{q}, \boldsymbol{\theta}_{10}, \ldots, \boldsymbol{\theta}_{\ell 0}, \boldsymbol{\theta}_{11}, \ldots, \boldsymbol{\theta}_{\ell 1}, \mu, \tau_0, \tau_1 | \mathbf{y}) \propto \prod_{i=1}^{\ell} \frac{\prod_{u=1}^{U_i} (\pi_{iu}^* \theta_{iu0})^{g_{iu0}}}{\left(\sum_{u=1}^{U_i} \pi_{iu}^* \theta_{iu0}\right)^{n_i - s_i}} \prod_{i=1}^{\ell} \frac{\prod_{u=1}^{U_i} (\pi_{iu}^* \theta_{iu1})^{g_{iu1}}}{\left(\sum_{u=1}^{U_i} \pi_{iu}^* \theta_{iu1}\right)^{s_i}}$$
$$\times \prod_{i=1}^{\ell} q_i^{s_i} (1 - q_i)^{n_i - s_i} \prod_{i=1}^{\ell} \frac{\prod_{u=1}^{U_i} \theta_{iu0}^{\theta_{iu0}^{(0)} \tau_0 - 1}}{D\left(\boldsymbol{\theta}_{i0}^{(0)} \tau_0\right)} \prod_{i=1}^{\ell} \frac{\prod_{u=1}^{U_i} \theta_{iu1}^{\theta_{iu1}^{(0)} \tau_0 - 1}}{D\left(\boldsymbol{\theta}_{i1}^{(0)} \tau_0\right)}$$
$$\times \frac{\prod_{i=1}^{\ell} q_i^{\mu\tau_1 - 1} (1 - q_i)^{(1-\mu)\tau_1 - 1}}{[B(\mu\tau_1, (1 - \mu)\tau_1)]^{\ell}} \frac{1}{(1 + \tau_0)^2} \frac{1}{(1 + \tau_1)^2}.$$

To improve the computations, we use the more optimal Rao-Blackwellization to get the posterior density of $\mathbf{q}$ (hence $\mathbf{p}$). Given the data, the joint posterior density can be expressed as

$$\pi(\mathbf{q}, \boldsymbol{\theta}, \mu, \tau_0, \tau_1 | \mathbf{y}) = \pi(\mathbf{q} | \boldsymbol{\theta}, \mu, \tau_0, \tau_1, \mathbf{y}) \pi(\boldsymbol{\theta}, \mu, \tau_0, \tau_1 | \mathbf{y}).$$

3.2 Computations

By integrating out $\mathbf{q}$ from the joint posterior of $\mathbf{q}, \boldsymbol{\theta}_{10}, \ldots, \boldsymbol{\theta}_{\ell 0}, \boldsymbol{\theta}_{11}, \ldots, \boldsymbol{\theta}_{\ell 1}, \mu, \tau_0, \tau_1$ given $\mathbf{y}$, we get the marginal joint posterior density of $\boldsymbol{\theta}_{10}, \ldots, \boldsymbol{\theta}_{\ell 0}, \boldsymbol{\theta}_{11}, \ldots, \boldsymbol{\theta}_{\ell 1}, \mu, \tau_0, \tau_1$ given $\mathbf{y}$,

$$\pi(\boldsymbol{\theta}_{10}, \ldots, \boldsymbol{\theta}_{\ell 0}, \boldsymbol{\theta}_{11}, \ldots, \boldsymbol{\theta}_{\ell 1}, \mu, \tau_0, \tau_1 | \mathbf{y}) \propto \frac{\prod_{i=1}^{\ell} B(s_i + \mu\tau_1, n_i - s_i + (1 - \mu)\tau_1)}{[B(\mu\tau_1, (1 - \mu)\tau_1)]^{\ell}}$$
$$\times \prod_{i=1}^{\ell} \frac{\prod_{u=1}^{U_i} \theta_{iu0}^{g_{iu0} + \theta_{iu0}^{(0)} \tau_0 - 1}}{\left(\sum_{u=1}^{U_i} \pi_{iu}^* \theta_{iu0}\right)^{n_i - s_i}} \prod_{i=1}^{\ell} \frac{\prod_{u=1}^{U_i} \theta_{iu1}^{g_{iu1} + \theta_{iu1}^{(0)} \tau_0 - 1}}{\left(\sum_{u=1}^{U_i} \pi_{iu}^* \theta_{iu1}\right)^{s_i}}$$
$$\times \frac{1}{\prod_{i=1}^{\ell} D\left(\boldsymbol{\theta}_{i0}^{(0)} \tau_0\right)} \frac{1}{\prod_{i=1}^{\ell} D\left(\boldsymbol{\theta}_{i1}^{(0)} \tau_0\right)} \frac{1}{(1 + \tau_0)^2} \frac{1}{(1 + \tau_1)^2}.$$

The conditional posterior density of $\mathbf{q}$ is given by

$$\pi(\mathbf{q} | \boldsymbol{\theta}_{10}, \ldots, \boldsymbol{\theta}_{\ell 0}, \boldsymbol{\theta}_{11}, \ldots, \boldsymbol{\theta}_{\ell 1}, \mu, \tau_0, \tau_1, \mathbf{y}) \propto \prod_{i=1}^{\ell} q_i^{s_i + \mu\tau_1 - 1} (1 - q_i)^{n_i - s_i + (1-\mu)\tau_1 - 1},$$

which we can sample directly.

We obtain a sample of $\mathbf{q}$ in the following manner: We draw each element of $(\boldsymbol{\theta}_{10}, \ldots, \boldsymbol{\theta}_{\ell 0}, \boldsymbol{\theta}_{11}, \ldots, \boldsymbol{\theta}_{\ell 1}, \mu, \tau_0, \tau_1)$ from the conditional posterior density of $\boldsymbol{\theta}_{10}, \ldots, \boldsymbol{\theta}_{\ell 0}, \boldsymbol{\theta}_{11}, \ldots, \boldsymbol{\theta}_{\ell 1}, \mu, \tau_0, \tau_1 | \mathbf{y}$ using the Metropolis-Hastings algorithm and a grid method, and then we draw $\mathbf{q}$ from the conditional posterior density of $\mathbf{q} | \boldsymbol{\theta}_{10}, \ldots, \boldsymbol{\theta}_{\ell 0}, \boldsymbol{\theta}_{11}, \ldots, \boldsymbol{\theta}_{\ell 1}, \mu, \tau_0, \tau_1, \mathbf{y}$.

We have monitored the convergence of the Metropolis-Hastings sampler using trace plots, autocorrelation plots and Geweke test of stationarity, which showed satisfactory performance.

The conditional posterior densities needed to execute the Metropolis-Hastings sampler are

$$\pi\left(\boldsymbol{\theta}_{i0} \,|\, \boldsymbol{\theta}_{j0},\ j \neq i,\ \boldsymbol{\theta}_{i1},\ i = 1, \ldots, \ell,\ \mu,\ \tau_0,\ \tau_1,\ \mathbf{y}\right) \propto \frac{\prod_{u=1}^{U_i} \theta_{iu0}^{g_{iu0} + \theta_{iu0}^{(0)}\tau_0 - 1}}{\left(\sum_{u=1}^{U_i} \pi_{iu}^{*} \theta_{iu0}\right)^{n_i - s_i}},\ i = 1, \ldots, \ell,$$

$$\pi\left(\boldsymbol{\theta}_{i1} \,|\, \boldsymbol{\theta}_{j1},\ j \neq i,\ \boldsymbol{\theta}_{i0},\ i = 1, \ldots, \ell,\ \mu,\ \tau_0,\ \tau_1,\ \mathbf{y}\right) \propto \frac{\prod_{u=1}^{U_i} \theta_{iu1}^{g_{iu1} + \theta_{iu1}^{(0)}\tau_0 - 1}}{\left(\sum_{u=1}^{U_i} \pi_{iu}^{*} \theta_{iu1}\right)^{s_i}},\ i = 1, \ldots, \ell,$$

$$\pi\left(\mu \,|\, \boldsymbol{\theta}_{10}, \ldots, \boldsymbol{\theta}_{\ell 0}, \boldsymbol{\theta}_{11}, \ldots, \boldsymbol{\theta}_{\ell 1},\ \tau_0,\ \tau_1,\ \mathbf{y}\right) \propto \frac{\prod_{i=1}^{\ell} B\left(s_i + \mu\tau_1,\ n_i - s_i + (1-\mu)\tau_1\right)}{\left[B\left(\mu\tau_1,\ (1-\mu)\tau_1\right)\right]^{\ell}},$$

and transforming τ_0 and τ_1 to respectively $\rho_0 = 1/(1+\tau_0)$ and $\rho_1 = 1/(1+\tau_1)$,

$$\pi\left(\rho_0 \,|\, \boldsymbol{\theta}_{10}, \ldots, \boldsymbol{\theta}_{\ell 0}, \boldsymbol{\theta}_{11}, \ldots, \boldsymbol{\theta}_{\ell 1},\ \mu,\ \tau_1,\ \mathbf{y}\right) \propto \left[\prod_{i=1}^{\ell} \frac{\prod_{u=1}^{U_i} \theta_{iu0}^{\theta_{iu0}^{(0)}\tau_0 - 1}}{D\left(\boldsymbol{\theta}_{i0}^{(0)}\tau_0\right)} \prod_{i=1}^{\ell} \frac{\prod_{u=1}^{U_i} \theta_{iu1}^{\theta_{iu1}^{(0)}\tau_0 - 1}}{D\left(\boldsymbol{\theta}_{i1}^{(0)}\tau_0\right)}\right]_{\tau_0 = (1-\rho_0)/\rho_0},$$

$$\pi\left(\rho_1 \,|\, \boldsymbol{\theta}_{10}, \ldots, \boldsymbol{\theta}_{\ell 0}, \boldsymbol{\theta}_{11}, \ldots, \boldsymbol{\theta}_{\ell 1},\ \mu,\ \rho_0,\ \mathbf{y}\right) \propto \left[\frac{\prod_{i=1}^{\ell} B\left(s_i + \mu\tau_1,\ n_i - s_i + (1-\mu)\tau_1\right)}{\left[B\left(\mu\tau_1,\ (1-\mu)\tau_1\right)\right]^{\ell}}\right]_{\tau_1 = (1-\rho_1)/\rho_1}.$$

In above formula, the $\boldsymbol{\theta}_{i0}$, $\boldsymbol{\theta}_{i1}$ and μ are conditionally independent.

We use Metropolis steps to sample $\boldsymbol{\theta}_{i0}$, $i = 1, \ldots, \ell$ and $\boldsymbol{\theta}_{i1}$, $i = 1, \ldots, \ell$ from conditional distributions $\pi\left(\boldsymbol{\theta}_{i0} \,|\, \boldsymbol{\theta}_{j0},\ j \neq i,\ \boldsymbol{\theta}_{i1},\ i = 1, \ldots, \ell,\ \mu,\ \tau_0,\ \tau_1,\ \mathbf{y}\right)$ and $\pi\left(\boldsymbol{\theta}_{i1} \,|\, \boldsymbol{\theta}_{j1},\ j \neq i,\ \boldsymbol{\theta}_{i0},\ i = 1, \ldots, \ell,\ \mu,\ \tau_0,\ \tau_1,\ \mathbf{y}\right)$, respectively. Let $\tau_{i0}^{*} = n_i - s_i + \tau_0$, $\theta_{iu0}^{*} = \left(g_{iu0} + \theta_{iu0}^{(0)}\tau_0\right)/\tau_{i0}^{*}$, $\tau_{i1}^{*} = s_i + \tau_0$, and $\theta_{iu1}^{*} = \left(g_{iu1} + \theta_{iu1}^{(0)}\tau_0\right)/\tau_{i1}^{*}$. For notational convenience, let $\boldsymbol{\theta}_{i0}^{*} = \left(\theta_{i10}^{*}, \ldots, \theta_{iU_i 0}^{*}\right)$ and $\boldsymbol{\theta}_{i1}^{*} = \left(\theta_{i11}^{*}, \ldots, \theta_{iU_i 1}^{*}\right)$. We choose overdispersed Dirichlet distributions as proposal densities for $\boldsymbol{\theta}_{i0}$ and $\boldsymbol{\theta}_{i1}$, $i = 1, \ldots, \ell$. In fact, the proposal densities are $\text{Dirichlet}\left(\boldsymbol{\theta}_{i0}^{*}\tilde{\tau}_{i0}\right)$ and $\text{Dirichlet}\left(\boldsymbol{\theta}_{i1}^{*}\tilde{\tau}_{i1}\right)$, where $\tilde{\tau}_{i0} = \tau_{i0}^{*}/\kappa_{i0}$ and $\tilde{\tau}_{i1} = \tau_{i1}^{*}/\kappa_{i1}$, for all $i = 1, \ldots, \ell$. Note that as the κ_{i0} and κ_{i1} increase, the dispersion tends to increase in the Dirichlet distribution.

Assuming that the Metropolis-Hastings sampler is at $\boldsymbol{\theta}_{i0}^{(r)}$, then the probability of accepting $\boldsymbol{\theta}_{i0}^{(r+1)}$ is

$$U_{r,r+1}^{(0)} = \min\left\{\frac{\psi\left(\boldsymbol{\theta}_{i0}^{(r+1)},\ \mathbf{y}\right)}{\psi\left(\boldsymbol{\theta}_{i0}^{(r)},\ \mathbf{y}\right)},\ 1\right\},$$

where $\psi\left(\boldsymbol{\theta}_{i0},\ \mathbf{y}\right) = \frac{\prod_{u=1}^{U_i} \theta_{iu0}^{\theta_{iu0}^{*}\tau_{i0}^{*}(1-1/\kappa_{i0})}}{\left(\sum_{u=1}^{U_i} \pi_{iu}^{*}\theta_{iu0}\right)^{n_i - s_i}}$, for all $i = 1, \ldots, \ell$. Also, if we assume that the Metropolis-Hastings sampler is at $\boldsymbol{\theta}_{i1}^{(r)}$, then the probability of accepting $\boldsymbol{\theta}_{i1}^{(r+1)}$ is

$$U_{r,r+1}^{(1)} = \min\left\{\frac{\psi\left(\boldsymbol{\theta}_{i1}^{(r+1)},\ \mathbf{y}\right)}{\psi\left(\boldsymbol{\theta}_{i1}^{(r)},\ \mathbf{y}\right)},\ 1\right\},$$

where $\psi\left(\boldsymbol{\theta}_{i1},\ \mathbf{y}\right) = \frac{\prod_{u=1}^{U_i} \theta_{iu1}^{\theta_{iu1}^{*}\tau_{i1}^{*}(1-1/\kappa_{i1})}}{\left(\sum_{u=1}^{U_i} \pi_{iu}^{*}\theta_{iu1}\right)^{s_i}}$, for all $i = 1, \ldots, \ell$.

The Metropolis step is obtained as follows: Assume the Markov chain is at $\boldsymbol{\theta}_{i0}^{(r)}$, a random vector $\boldsymbol{\theta}_{i0}^{(r+1)}$ is drawn from the proposal density with properly chosen κ_{i0}, and $U_{r,r+1}^{(0)}$ is computed. A random

uniform deviate U in $[0, 1]$ is drawn, and if $U \le U^{(0)}_{r,r+1}$, then random vector $\boldsymbol{\theta}^{(r+1)}_{i0}$ is accepted, otherwise the chain stays at $\boldsymbol{\theta}^{(r)}_{i0}$. This algorithm is applied for all $i = 1, \ldots, \ell$. The Metropolis step is utilized in a manner similar to that for $\boldsymbol{\theta}_{i1}$ for all $i = 1, \ldots, \ell$.

We generate μ, ρ_0 and ρ_1, using the grid method; see CNK. Once we get the sample from the posterior $\pi(\mathbf{q}, \boldsymbol{\theta}_{10}, \ldots, \boldsymbol{\theta}_{\ell 0}, \boldsymbol{\theta}_{11}, \ldots, \boldsymbol{\theta}_{\ell 1}, \mu, \tau_0, \tau_1 | \mathbf{y})$, by retransforming from q_i to p_i,

$$p_i = \frac{a_{i0} q_i}{a_{i0} q + a_{i1}(1 - q_i)},$$

we can draw a sample from the posterior distribution of $\pi(\mathbf{p}, \boldsymbol{\theta}_{10}, \ldots, \boldsymbol{\theta}_{\ell 0}, \boldsymbol{\theta}_{11}, \ldots, \boldsymbol{\theta}_{\ell 1}, \mu, \tau_0, \tau_1 | \mathbf{y})$.

Once $\mathbf{p}$ is estimated, we draw the entire finite population values, $y_{i1}, \ldots, y_{iN_i}$, independently from Bernoulli(p_i), $i = 1, \ldots, \ell$. This is surrogate sampling (e.g., Nandram, 2007). So we have corrected the observed biased sample and replaced it by a surrogate sample for $\mathbf{p}$ that we obtained from the heterogeneous nonignorable selection model. We can obtain a sample of P_i by drawing $\sum_{i=1}^{N_i} y_{ij}$ from Binomial(N_i, p_i) and by dividing the result by N_i for all $i = 1, \ldots, \ell$.

The selection mechanism is similar to the missing data mechanism. So it is possible to incorporate missing data into our framework, or independently (i.e., on its own) we can assume that the missing data are "missing not at random" and a nonignorable nonresponse model can be used to adjust a population model, see Nandram and Choi (2010).

4. Numerical studies

In Section 4.1, we describe an example on severe activity limitation. We present the results of the IS, HoS and HeS models for the comparison. In Section 4.2, we describe a simulation study to assess the performance of the three models under two kinds of assumptions of the distribution of the sample selection probabilities. That is, data are generated from either the homogeneous nonignorable or the heterogeneous nonignorable selection model and all three models are fit.

4.1 Illustrative example

In our application, we use data from the 1995 National Health Interview Survey (NHIS95). These data were first used by Nandram, Bai and Choi (2011) to estimate change point in activity limitation. NBBS constructed synthetic data arising from a segment of NHIS95 for this study. For adults, 30-80 years old, NBBS analyzed data on severe activity limitation (SAL) for a single area (no pooling), where $y = 1$ if an adult has SAL and $y = 0$ otherwise. CNK used the data from NBBS to perform small area estimation on SAL. We use the data in a manner similar to NBBS but we fit a more general model, the key contribution of this paper.

NBBS formed twelve domains (small areas) by crossing education, sex and race. They have categorized education into three levels (pre-college: L, college: M and post-college: H). Sex (male: M and

female: F) and Race (white: W and nonwhite: B) have two levels each. We will continue to call these domains LMW, LMB, LFW, LFB, MMW, MMB, MFW, MFB, HMW, HMB, HFW, HFB (e.g., LMW: white males with pre-college education, LMB: black males with pre-college education, etc.).

NHIS95 used a multistage sample design to draw samples from the population of the United States. Therefore, it is necessary to use an adult's survey weight for accurate analysis. See NBBS and CNK for a discussion of the survey weights. Like NBBS and CNK, we considered the reciprocal of a survey weight as the "selection" probability of each adult. Selection probabilities are a major part of the survey weights. It would have been better if we had the selection probabilities. But this is an approximation we make in our data analysis. For convenience, we have presented the data again in Table 4.1.

Table 4.1
Summaries of the data on severe activity limitation (SAL) including selection probabilities

Domain	n	s	p	CV	$f = n/N$	avg0	avg1	p -value
LMW	174	26	0.149	0.181	0.0000303	0.0004059	0.0003706	0.025*
LMB	31	10	0.323	0.260	0.0000280	0.0003179	0.0003167	0.614
LFW	200	22	0.110	0.201	0.0000325	0.0004523	0.0004381	1.00
LFB	40	9	0.225	0.293	0.0000286	0.0003152	0.0003740	0.910
MMW	760	56	0.074	0.129	0.0000227	0.0002710	0.0002901	0.006*
MMB	151	18	0.119	0.221	0.0000250	0.0002820	0.0002847	0.056
MFW	892	42	0.047	0.151	0.0000233	0.0002891	0.0002415	0.837
MFB	200	21	0.105	0.206	0.0000278	0.0003064	0.0003377	0.469
HMW	756	14	0.019	0.265	0.0000213	0.0002484	0.0001919	0.095
HMB	124	7	0.056	0.367	0.0000238	0.0002702	0.0003506	0.146
HFW	779	22	0.028	0.210	0.0000219	0.0002575	0.0002976	0.072
HFB	168	2	0.012	0.703	0.0000257	0.0002969	0.0001948	1.00

NOTE: Here n is the total sample size, s is the number of adults with SAL, and $p = s/n$; $f = n/N$ is the sampling fraction; CV = $\frac{\sqrt{(1-p)}}{\sqrt{np}}$ is the coefficient of variation; avg0 is the average of the selection probabilities for $y = 0$ (SAL, no) and avg1 is the average of the selection probabilities for $y = 1$ (SAL, yes); p -value corresponds to that of a chi-squared test of equality of $\theta_{iu0} = \theta_{iu1}$, $u = 1, \ldots, U_i$ $(U_i = 5)$; domains are formed by crossing education (L, M, H), sex (M, F) and race (W, B).

We reduce the sample size from the original data set to increase the effect of small area model. For the HeS model, we order the selection probabilities from smallest to largest $\pi_{(1)}, \ldots, \pi_{(n_i)}$ within each small area. Let the quantiles be $t_{i1} = \pi_{(0.20n_i)}$, $t_{i2} = \pi_{(0.40n_i)}$, $t_{i3} = \pi_{(0.60n_i)}$, $t_{i4} = \pi_{(0.80n_i)}$ and let $t_{i0} = \pi_{(1)}$ and $t_{i5} = \pi_{(n_i)}$ for $i = 1, \ldots, \ell$. We define $\pi^*_{iu} = (t_{i(u-1)} + t_{iu})/2$, $u = 1, \ldots, U_i$, $i = 1, \ldots, \ell$ (i.e., the mid point of each quantile within each small area). Note that θ_{iuy} is the proportion of sampled units in the u^{th} quantile conditional on $y = 0, 1$ for $i = 1, \ldots, \ell$. If the θ_{iu0} are considerably different from θ_{iu1} in some areas, there is strong evidence that the sampled values are biased. To specify $\boldsymbol{\theta}^{(0)}_{i0}$ and $\boldsymbol{\theta}^{(0)}_{i1}$ in the prior distributions, we take $\boldsymbol{\theta}^{(0)}_{iy} = \hat{\boldsymbol{\theta}}_{iy}$, the maximum likelihood estimator of $\boldsymbol{\theta}_{iy}$, $y = 0, 1$, $i = 1, \ldots, \ell$, which we call a "MLE" prior (see NBBS).

For our data, mixing is slow, so we draw 120,000 samples and burn in 60,000. Then we take every 30^{th} iterate to obtain a sample of 2,000 iterates for inference. This burn-in period is sufficiently long to get random samples, which is based on the trace plots and Geweke test. We have enough samples since the effective sample size (ESS) of parameters sampled from an MCMC should be less than or equal to 2,000. The correlation is nonsignificant for all parameters. Also, stationarity of our sampler is

demonstrated by Geweke test. The results of convergence diagnostic for hyper-parameters are shown in Table 4.8 and Figure 4.2. By taking $\boldsymbol{\kappa}_0 = (7.3, 3, 11, 4.8, 8.5, 9, 11.5, 10, 10, 9.7, 10.2, 11)$ and $\boldsymbol{\kappa}_1 = (3.2, 1.7, 2.2, 1.5, 5, 2.2, 3.3, 2.2, 2.5, 1.5, 1.5, 1.3)$, we obtained acceptance rates between 25% and 45% of Metropolis-Hastings samplers.

The summaries of data are shown in Table 4.1. The averages of the selection probabilities in each area are mostly similar for $y = 0$ and $y = 1$. Since the sample sizes within the domains are not quite large, the sampling fractions are very small. Thus, the simultaneous analysis using a small area model is appropriate. In column 4 of Table 4.1 we have also presented the proportion of individuals with SAL, and we can see that this value is relatively large for low level education. We compared the counts in the two sets of bins from the histograms of the selection probabilities for $y = 0, 1$ in each domain. In fact, this is a test of independence in a $2 \times U_i$ categorical table, and we use a chi-squared test and a Fisher's exact test for equality of $\theta_{iu0} = \theta_{iu1}$, $u = 1, \ldots, U_i$ ($U_i = 5$ cells) in each domain. As is evident from the p-values which are presented in the last column of Table 4.1, selection bias should matter mostly in Domain MMW and perhaps in Domains LMW, MMB, HMW and HFW. As a measure of selection bias, we also calculated the biserial correlation between the y_{ij} and π_{ij} for all areas combined and got a value of 0.033 with a p-value of 0.03.

In Table 4.2, we provide the results of the unpooled (individual analysis) of the finite population proportions under the ignorable selection model and the nonignorable selection model for each domain separately (the nonignorable selection model is the NBBS model). We compare them using the posterior means (PM), the posterior standard deviations (PSD) of PMs, the coefficient of variation (CV) and 95% highest posterior density (HPD) intervals. This is a repetition of the analysis under the NBBS model.

In Table 4.3, we compare summaries of the pooled estimators of the finite population proportions under the IS model, the HoS model and the HeS model for the 12 domains. The effect of the selection bias is seen because the PMs under the IS model and the other models are different. The estimators under the HoS model and the HeS model are similar in some domains because the selection bias is more severe for some domains. The PSDs under the HoS model and the HeS model are bigger than under the IS model in most domains, but the 95% HPD intervals overlap. The effect of the simultaneous analysis is also seen because the PMs for pooled estimators are smoothed relative to PMs for separately analyzed finite populations.

The posterior summaries of $\boldsymbol{\theta}_{i0}$ and $\boldsymbol{\theta}_{i1}$ for $i = 1, \ldots, \ell$ are shown in Tables 4.4. The θ_{iu0} are different from the θ_{iu1} in some areas. It is strong evidence to indicate the presence of selection bias. We also present the posterior summaries of $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ under the HoS model in Table 4.5. Note that in Tables 4.4-4.5, these θ_{i0} and θ_{i1} are not selection probabilities. The $\boldsymbol{\theta}$ under the HoS model appears different from those under the HeS model. We present the posterior summaries of a_{i0} and a_{i1} for $i = 1, \ldots, \ell$ in Table 4.7, and a_0 and a_1 of the HoS model in Table 4.6. These values are very small, but its ratio is large in each area. It is also strong evidence of selection bias, albeit these are small sample sizes within areas.

Table 4.2
Comparisons of the unpooled estimators of the finite population proportions under the ignorable and nonignorable selection models for individual domain

Domain	Ignorable Selection Model				Nonignorable Selection Model			
	PM	PSD	CV	CI	PM	PSD	CV	CI
LMW	0.154	0.027	0.175	(0.098, 0.214)	0.188	0.033	0.176	(0.123, 0.265)
LMB	0.333	0.080	0.240	(0.177, 0.525)	0.336	0.083	0.247	(0.170, 0.531)
LFW	0.114	0.022	0.193	(0.068, 0.165)	0.113	0.022	0.195	(0.069, 0.165)
LFB	0.244	0.066	0.270	(0.108, 0.394)	0.215	0.064	0.298	(0.090, 0.361)
MMW	0.075	0.010	0.133	(0.055, 0.100)	0.068	0.009	0.132	(0.050, 0.086)
MMB	0.124	0.026	0.210	(0.073, 0.185)	0.114	0.024	0.211	(0.065, 0.171)
MFW	0.048	0.008	0.167	(0.032, 0.065)	0.052	0.008	0.154	(0.036, 0.072)
MFB	0.108	0.022	0.204	(0.066, 0.161)	0.104	0.021	0.202	(0.060, 0.148)
HMW	0.020	0.005	0.250	(0.008, 0.032)	0.021	0.005	0.238	(0.011, 0.034)
HMB	0.065	0.022	0.338	(0.021, 0.119)	0.046	0.018	0.391	(0.014, 0.089)
HFW	0.030	0.006	0.200	(0.017, 0.045)	0.021	0.005	0.238	(0.011, 0.032)
HFB	0.017	0.010	0.588	(0.001, 0.043)	0.021	0.012	0.571	(0.001, 0.048)

NOTE: Here PM is the posterior mean; PSD is the posterior standard deviation; CV is the coefficient of variation; CI is the 95% HPD interval.

Table 4.3
Comparisons of the pooled estimators of the finite population proportions under the ignorable selection (IS) model, homogeneous nonignorable selection (HoS) model and heterogeneous nonignorable selection (HeS) models by domain

Domain	Model	PM	PSD	CV	CI
LMW	IS	0.148	0.026	0.176	(0.094, 0.210)
	HoS	0.177	0.030	0.169	(0.123, 0.238)
	HeS	0.172	0.037	0.215	(0.099, 0.264)
LMB	IS	0.264	0.071	0.269	(0.109, 0.426)
	HoS	0.310	0.082	0.265	(0.167, 0.483)
	HeS	0.267	0.076	0.285	(0.120, 0.449)
LFW	IS	0.110	0.021	0.191	(0.062, 0.158)
	HoS	0.134	0.026	0.194	(0.088, 0.189)
	HeS	0.116	0.027	0.233	(0.058, 0.179)
LFB	IS	0.198	0.057	0.288	(0.083, 0.326)
	HoS	0.237	0.064	0.270	(0.122, 0.377)
	HeS	0.192	0.060	0.313	(0.074, 0.332)
MMW	IS	0.074	0.009	0.122	(0.054, 0.095)
	HoS	0.091	0.011	0.121	(0.071, 0.115)
	HeS	0.081	0.013	0.160	(0.054, 0.111)
MMB	IS	0.119	0.026	0.218	(0.067, 0.178)
	HoS	0.144	0.030	0.208	(0.092, 0.207)
	HeS	0.120	0.027	0.225	(0.060, 0.183)
MFW	IS	0.048	0.007	0.146	(0.032, 0.064)
	HoS	0.059	0.009	0.153	(0.044, 0.078)
	HeS	0.060	0.011	0.183	(0.036, 0.085)
MFB	IS	0.106	0.021	0.198	(0.062, 0.154)
	HoS	0.128	0.026	0.203	(0.084, 0.183)
	HeS	0.105	0.023	0.219	(0.059, 0.160)
HMW	IS	0.020	0.005	0.250	(0.010, 0.032)
	HoS	0.025	0.006	0.240	(0.014, 0.039)
	HeS	0.027	0.008	0.296	(0.013, 0.047)
HMB	IS	0.062	0.021	0.339	(0.019, 0.111)
	HoS	0.075	0.025	0.333	(0.035, 0.131)
	HeS	0.057	0.020	0.351	(0.019, 0.108)
HFW	IS	0.029	0.006	0.207	(0.017, 0.044)
	HoS	0.037	0.008	0.216	(0.023, 0.053)
	HeS	0.027	0.006	0.222	(0.014, 0.042)
HFB	IS	0.018	0.010	0.556	(0.001, 0.043)
	HoS	0.022	0.012	0.545	(0.005, 0.050)
	HeS	0.020	0.012	0.600	(0.001, 0.048)

Table 4.4
Comparison of θ_{i0} and θ_{i1} using PM, PSD, NSE and 95% HPD interval under the heterogeneous nonignorable selection model

Domain	u	θ_0				θ_1			
		PM	PSD	CV	CI	PM	PSD	CV	CI
LMW	1	0.260	0.083	0.319	(0.096, 0.447)	0.477	0.109	0.229	(0.252, 0.714)
	2	0.213	0.075	0.352	(0.073, 0.397)	0.195	0.087	0.446	(0.048, 0.401)
	3	0.214	0.070	0.327	(0.084, 0.383)	0.126	0.070	0.556	(0.008, 0.296)
	4	0.189	0.064	0.339	(0.065, 0.342)	0.079	0.052	0.658	(0.003, 0.208)
	5	0.124	0.048	0.387	(0.027, 0.235)	0.121	0.057	0.471	(0.019, 0.252)
LMB	1	0.318	0.098	0.308	(0.101, 0.536)	0.355	0.100	0.282	(0.140, 0.570)
	2	0.279	0.086	0.308	(0.058, 0.422)	0.176	0.080	0.455	(0.020, 0.361)
	3	0.209	0.083	0.397	(0.048, 0.340)	0.148	0.073	0.493	(0.021, 0.324)
	4	0.126	0.065	0.516	(0.015, 0.286)	0.229	0.083	0.362	(0.063, 0.414)
	5	0.114	0.059	0.518	(0.012, 0.254)	0.093	0.054	0.581	(0.007, 0.224)
LFW	1	0.279	0.087	0.312	(0.101, 0.473)	0.301	0.090	0.299	(0.113, 0.500)
	2	0.232	0.083	0.358	(0.075, 0.431)	0.215	0.083	0.386	(0.051, 0.410)
	3	0.194	0.075	0.387	(0.054, 0.371)	0.203	0.079	0.389	(0.056, 0.390)
	4	0.167	0.065	0.389	(0.046, 0.325)	0.191	0.071	0.372	(0.046, 0.348)
	5	0.128	0.052	0.406	(0.025, 0.245)	0.089	0.045	0.506	(0.008, 0.198)
LFB	1	0.308	0.098	0.318	(0.120, 0.538)	0.276	0.094	0.341	(0.084, 0.480)
	2	0.247	0.095	0.385	(0.058, 0.463)	0.183	0.083	0.454	(0.024, 0.377)
	3	0.192	0.085	0.443	(0.032, 0.391)	0.200	0.086	0.430	(0.044, 0.410)
	4	0.158	0.074	0.468	(0.023, 0.326)	0.202	0.081	0.401	(0.050, 0.392)
	5	0.095	0.058	0.611	(0.005, 0.239)	0.139	0.069	0.496	(0.024, 0.308)
MMW	1	0.204	0.043	0.211	(0.116, 0.304)	0.167	0.078	0.467	(0.035, 0.355)
	2	0.221	0.044	0.199	(0.133, 0.326)	0.186	0.081	0.435	(0.039, 0.380)
	3	0.213	0.043	0.202	(0.126, 0.314)	0.198	0.082	0.414	(0.043, 0.396)
	4	0.189	0.040	0.212	(0.110, 0.282)	0.359	0.094	0.262	(0.174, 0.594)
	5	0.173	0.033	0.191	(0.104, 0.248)	0.090	0.047	0.522	(0.013, 0.205)
MMB	1	0.284	0.086	0.303	(0.107, 0.472)	0.248	0.095	0.383	(0.059, 0.458)
	2	0.230	0.079	0.343	(0.088, 0.418)	0.168	0.078	0.464	(0.023, 0.347)
	3	0.196	0.068	0.347	(0.055, 0.351)	0.176	0.078	0.443	(0.034, 0.353)
	4	0.148	0.062	0.419	(0.034, 0.297)	0.339	0.096	0.283	(0.134, 0.552)
	5	0.142	0.053	0.373	(0.038, 0.264)	0.069	0.046	0.667	(0.002, 0.186)
MFW	1	0.212	0.045	0.212	(0.127, 0.323)	0.305	0.093	0.305	(0.110, 0.520)
	2	0.218	0.045	0.206	(0.122, 0.321)	0.248	0.085	0.343	(0.084, 0.446)
	3	0.207	0.043	0.208	(0.117, 0.309)	0.191	0.076	0.398	(0.042, 0.364)
	4	0.201	0.043	0.214	(0.117, 0.311)	0.186	0.073	0.392	(0.049, 0.364)
	5	0.162	0.035	0.216	(0.090, 0.246)	0.071	0.038	0.535	(0.006, 0.161)
MFB	1	0.260	0.077	0.296	(0.110, 0.427)	0.331	0.093	0.281	(0.155, 0.568)
	2	0.229	0.073	0.319	(0.071, 0.387)	0.155	0.070	0.452	(0.029, 0.327)
	3	0.203	0.068	0.335	(0.067, 0.356)	0.156	0.069	0.442	(0.032, 0.319)
	4	0.179	0.061	0.341	(0.058, 0.317)	0.189	0.072	0.381	(0.050, 0.362)
	5	0.129	0.049	0.380	(0.032, 0.242)	0.168	0.064	0.381	(0.039, 0.319)
HMW	1	0.224	0.045	0.201	(0.138, 0.333)	0.277	0.105	0.379	(0.079, 0.531)
	2	0.206	0.044	0.214	(0.119, 0.313)	0.282	0.104	0.369	(0.074, 0.513)
	3	0.193	0.044	0.228	(0.104, 0.295)	0.257	0.099	0.385	(0.071, 0.491)
	4	0.210	0.044	0.210	(0.121, 0.312)	0.162	0.079	0.488	(0.025, 0.350)
	5	0.166	0.034	0.205	(0.099, 0.247)	0.022	0.026	1.182	(0.000, 0.097)
HMB	1	0.296	0.097	0.328	(0.095, 0.504)	0.265	0.099	0.374	(0.072, 0.493)
	2	0.223	0.085	0.381	(0.062, 0.423)	0.203	0.091	0.448	(0.027, 0.417)
	3	0.204	0.078	0.382	(0.058, 0.391)	0.125	0.073	0.584	(0.004, 0.305)
	4	0.166	0.070	0.422	(0.040, 0.323)	0.266	0.091	0.342	(0.091, 0.467)
	5	0.112	0.050	0.446	(0.026, 0.236)	0.142	0.069	0.486	(0.021, 0.307)
HFW	1	0.222	0.045	0.203	(0.126, 0.322)	0.191	0.078	0.408	(0.038, 0.367)
	2	0.217	0.047	0.217	(0.125, 0.329)	0.184	0.073	0.397	(0.039, 0.356)
	3	0.207	0.044	0.213	(0.110, 0.304)	0.169	0.071	0.420	(0.030, 0.332)
	4	0.199	0.043	0.216	(0.106, 0.296)	0.260	0.081	0.312	(0.097, 0.442)
	5	0.155	0.034	0.219	(0.079, 0.229)	0.196	0.050	0.255	(0.081, 0.348)
HFB	1	0.274	0.084	0.307	(0.116, 0.470)	0.366	0.114	0.311	(0.136, 0.631)
	2	0.223	0.075	0.336	(0.072, 0.400)	0.189	0.093	0.492	(0.012, 0.400)
	3	0.198	0.071	0.359	(0.060, 0.364)	0.212	0.095	0.448	(0.018, 0.421)
	4	0.177	0.064	0.362	(0.049, 0.317)	0.178	0.088	0.494	(0.024, 0.384)
	5	0.128	0.050	0.391	(0.035, 0.246)	0.055	0.050	0.909	(0.000, 0.194)

NOTE: u indicates the five intervals for the selection probabilities.

Table 4.5
Comparison of θ_0 and θ_1 using PM, PSD, NSE and 95% CI under the homogeneous nonignorable selection model

u	θ_0				θ_1			
	PM	PSD	CV	CI	PM	PSD	CV	CI
1	0.384	0.011	0.029	(0.363, 0.405)	0.632	0.036	0.057	(0.559, 0.700)
2	0.222	0.008	0.036	(0.206, 0.237)	0.160	0.026	0.163	(0.114, 0.217)
3	0.189	0.007	0.037	(0.176, 0.203)	0.112	0.018	0.161	(0.080, 0.150)
4	0.153	0.006	0.039	(0.142, 0.164)	0.060	0.009	0.150	(0.045, 0.079)
5	0.053	0.002	0.038	(0.049, 0.057)	0.036	0.005	0.139	(0.027, 0.047)

NOTE: u indicates the five intervals for the selection probabilities.

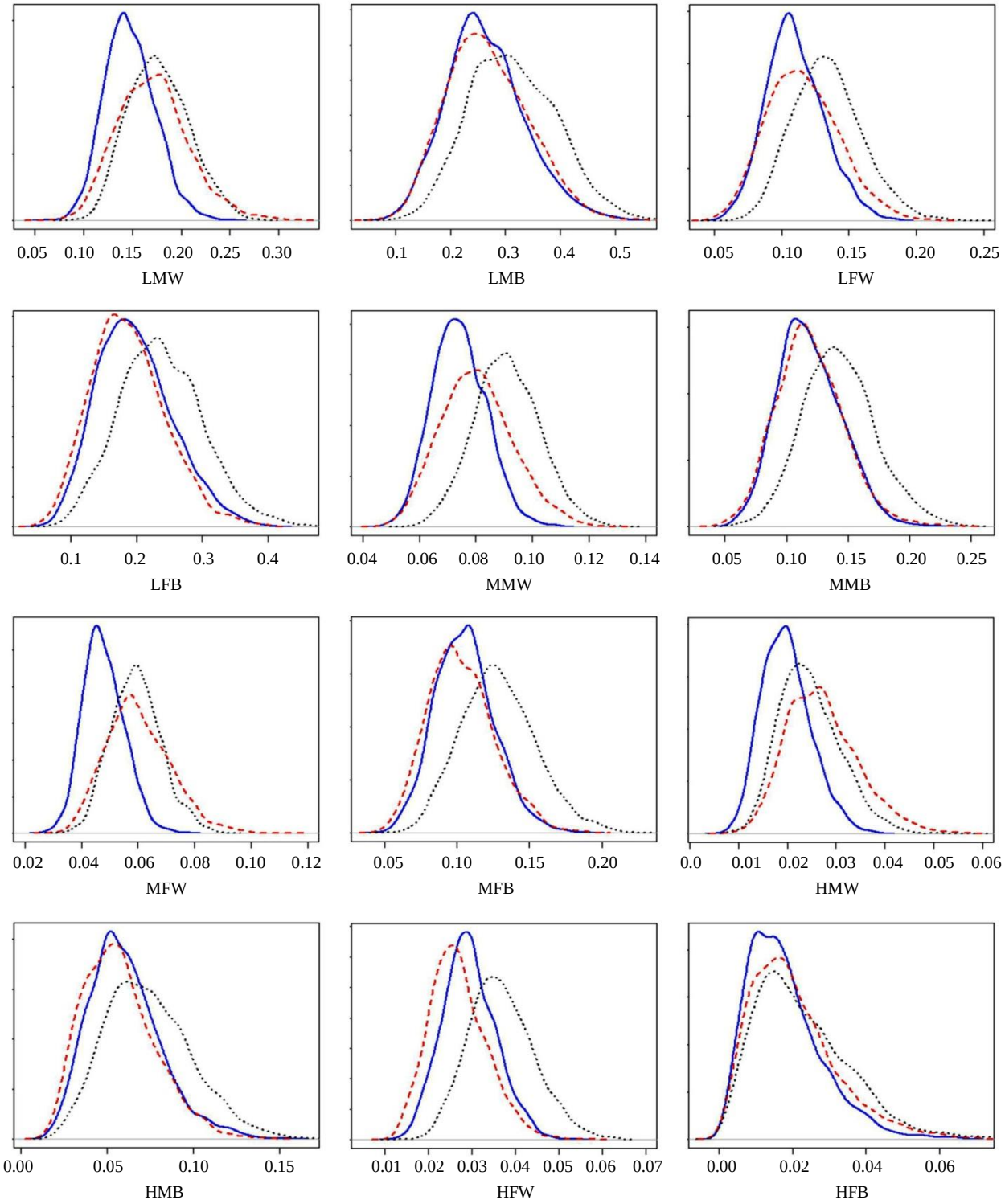
Table 4.6
Comparison of a_0 and a_1 using PM and PSD and NSE and 95% CI under the homogeneous nonignorable selection model

a_y	PM	PSD	CV	CI
a_0	0.000222	0.000002	0.009009	(0.000218, 0.000227)
a_1	0.000177	0.000006	0.033898	(0.000164, 0.000191)

Table 4.7
Comparison of a_{i0} and a_{i1} using PM, PSD, NSE and 95% HPD interval under the heterogeneous nonignorable selection model for each area

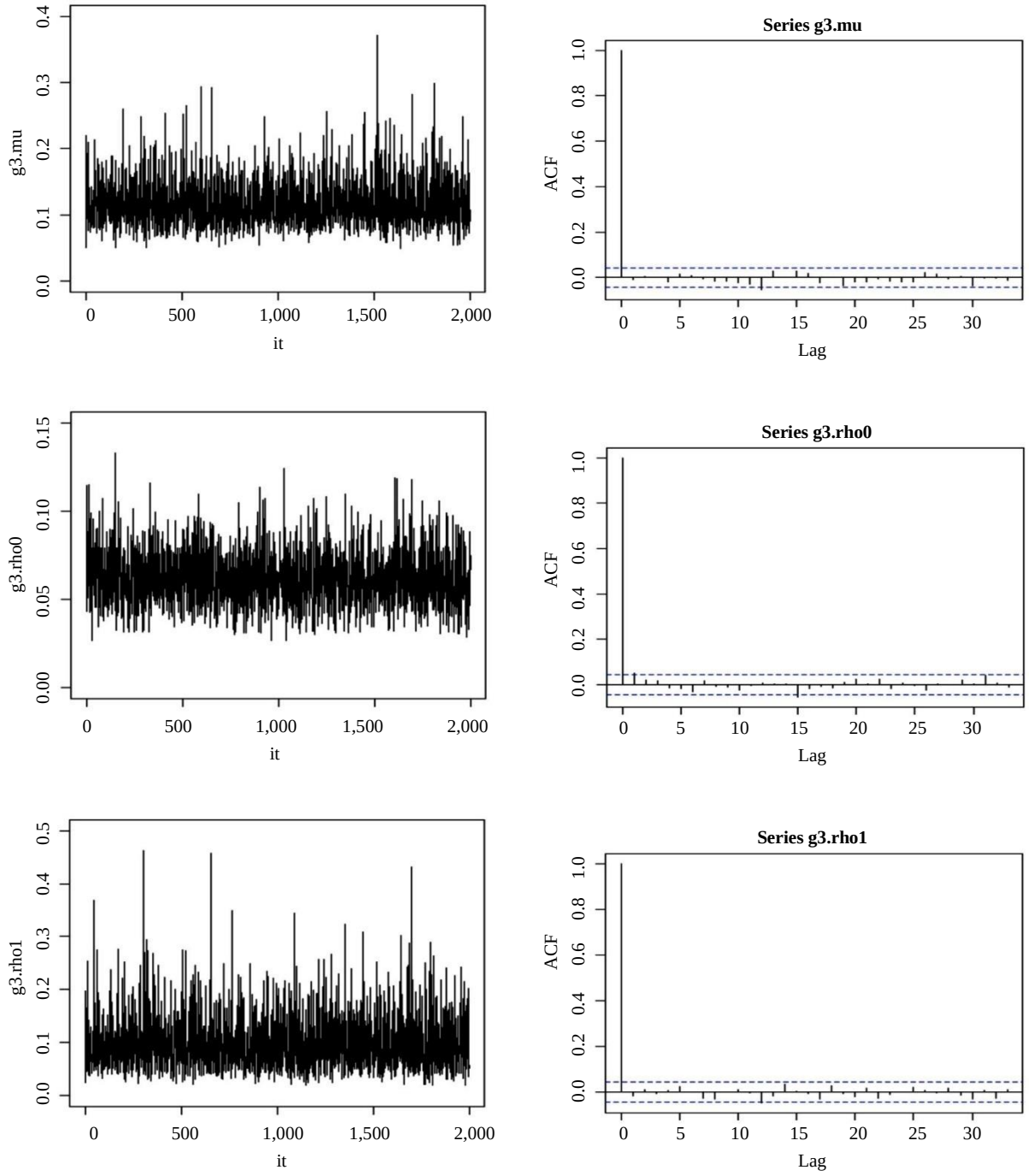
Domain	a_{iy}	PM	PSD	CV	CI
LMW	a_0	0.000367	0.000034	0.092643	(0.000296, 0.000441)
	a_1	0.000307	0.000040	0.130293	(0.000225, 0.000400)
LMB	a_0	0.000283	0.000021	0.074205	(0.000241, 0.000332)
	a_1	0.000282	0.000020	0.070922	(0.000240, 0.000331)
LFW	a_0	0.000411	0.000042	0.102190	(0.000317, 0.000504)
	a_1	0.000391	0.000040	0.102302	(0.000307, 0.000489)
LFB	a_0	0.000294	0.000022	0.074830	(0.000247, 0.000342)
	a_1	0.000314	0.000023	0.073248	(0.000268, 0.000369)
MMW	a_0	0.000286	0.000017	0.059441	(0.000248, 0.000323)
	a_1	0.000264	0.000025	0.094697	(0.000209, 0.000327)
MMB	a_0	0.000262	0.000018	0.068702	(0.000223, 0.000304)
	a_1	0.000260	0.000018	0.069231	(0.000224, 0.000304)
MFW	a_0	0.000313	0.000023	0.073482	(0.000263, 0.000362)
	a_1	0.000247	0.000026	0.105263	(0.000194, 0.000312)
MFB	a_0	0.000289	0.000017	0.058824	(0.000253, 0.000332)
	a_1	0.000294	0.000022	0.074830	(0.000246, 0.000345)
HMW	a_0	0.000263	0.000016	0.060837	(0.000228, 0.000298)
	a_1	0.000192	0.000016	0.083333	(0.000157, 0.000234)
HMB	a_0	0.000250	0.000022	0.088000	(0.000201, 0.000299)
	a_1	0.000270	0.000027	0.100000	(0.000212, 0.000335)
HFW	a_0	0.000293	0.000021	0.071672	(0.000246, 0.000340)
	a_1	0.000324	0.000037	0.114198	(0.000248, 0.000410)
HFB	a_0	0.000278	0.000021	0.075540	(0.000234, 0.000324)
	a_1	0.000246	0.000024	0.097561	(0.000196, 0.000303)

Figure 4.1 The posterior densities of the finite population proportions under the three models for the SAL data with 12 areas.



NOTE: Here the solid line represents the densities of P under the IS model, the dotted line represents that under the HoS model, and the dashed line represents that under the HeS model.

Figure 4.2 Trace plot and Autocorrelation plot for μ, ρ_0, ρ_1.



We present the posterior densities of the finite population proportions in Figure 4.1. The solid line represents the density of P under the IS model, the dotted line represents that under the HoS model, and the dashed line represents that under the HeS model for each domain. We can see that the plots of IS model are shifted due to the effect of the selection bias.

Table 4.8
Geweke convergence diagnostic (Z-score) and effective sample size (ESS) for μ, ρ_0, ρ_1

parameter	Geweke's statistic	p -value	ESS
μ	-0.4549	0.6492	2,000
ρ_0	1.7750	0.0759	1,809
ρ_1	-1.0331	0.3015	2,000

Our example shows that it is important to include a component for the selection mechanism into a model when a biased sample is available, otherwise there is likely to be misleading estimates of the small area proportions. Because the distribution of the sample selection probabilities is a little different across domains, the heterogeneity assumption is more reasonable for our numerical example.

4.2 Simulation study

We perform a simulation study to assess the accuracy of our model. Specifically, we consider two situations, which are homogeneous or heterogeneous distributions of the sample selection probabilities to show how different the IS, HoS and HeS models can be. We also show that when the correlation between the binary responses and the selection probabilities is strong, the IS model can perform badly.

To perform a simulation study, we generate $\ell = 12$ finite populations with i^{th} finite population having $N_i = 1{,}000$ units and selection probability π_{ij}, $j = 1, \ldots, 1{,}000$, $i = 1, \ldots, 12$. Then samples of size $n_i = 50$, $i = 1, \ldots, 12$ are taken from each area. We have generated 100 data sets. The outline of the data collection is as follows.

Step 1. Generate $p_i \sim U(0.2, 0.7)$, $i = 1, \ldots, \ell$. These values are true small population proportions.

Step 2. Set $\tau = 100$, $f_i = n_i / N_i$, $a = 0.975$, $\mu_0 = af$, $\mu_1 = f/a$, $i = 1, \ldots, \ell$.

Step 3. Generate $u \sim U(0, 1)$. If $u \leq p_i$ then we set $y_{ij} = 1$; otherwise set $y_{ij} = 0$, $j = 1, \ldots, N_i$, $i = 1, \ldots, \ell$.

Step 4. If $y_{ij} = 1$ then we generate $\pi_{ij} \sim \text{Beta}(\mu_1\tau, (1-\mu_1)\tau)$; if $y_{ij} = 0$ then we generate $\pi_{ij} \sim \text{Beta}(\mu_0\tau, (1-\mu_0)\tau)$ for $j = 1, \ldots, N_i$, $i = 1, \ldots, \ell$.

Step 5. Sample n_i units by systematic PPS sampling with probabilities $\frac{n_{ij}\pi_{ij}}{\sum_{i=1}^{N_i}\pi_{ij}}$.

We control the biserial correlation between the binary responses and the selection probabilities by changing the a values in Step 2. For example, if we set $a = 0.975, 0.95, 0.9, 0.8, 0.7$ in the simulation scheme, then the biserial correlations are respectably $\rho = 0.05, 0.1, 0.2, 0.4, 0.7$. We generate data having several biserial correlations, and categorized them as three levels (Low: $\rho < 0.3$, Medium: $0.3 \leq \rho < 0.6$, High: $\rho \geq 0.6$). These correspond to weak, medium and strong selection bias.

We can also control the heterogeneity or homogeneity assumptions by changing a support of the sample selection probabilities. For homogeneity assumption, we generate π_{ij} of Step 4 from similar

supports for all areas. On the other hand, we set up a support of π_{ij} from a different interval by area to make heterogeneous distributions of the sample selection probabilities.

To compare the performance of three models, we compute several frequentist measures. First, we calculate the finite population proportion $P_i^{(h)}$, the posterior mean $\mathrm{PM}_i^{(h)}$ and the posterior standard deviation $\mathrm{PSD}_i^{(h)}$, $i = 1, \ldots, \ell$, $h = 1, \ldots, 100$. Then we compute the absolute bias $\mathrm{AB}_i^{(h)} = \left| \mathrm{PM}_i^{(h)} - P_i^{(h)} \right|$ and the root mean squared error $\mathrm{RMSE}_i^{(h)} = \sqrt{\mathrm{PSD}_i^{(h)2} + \mathrm{AB}_i^{(h)2}}$, $i = 1, \ldots, \ell$, $h = 1, \ldots, 100$. Using these frequentist quantities we obtain $\mathrm{AB}_i = \frac{1}{100}\sum_{h=1}^{100} \mathrm{AB}_i^{(h)}$ and $\mathrm{RMSE}_i = \frac{1}{100}\sum_{h=1}^{100} \mathrm{RMSE}_i^{(h)}$. Also we compute the 95% highest posterior density (HPD) interval for each of the 100 simulated runs in each area. Then we look at the width $\left(W_i^{(h)}\right)$ and the HPD incidence $\left(I_i^{(h)}\right)$. Let $I_i^{(h)} = 1$ if the 95% HPD interval contains the true value P_i and let $I_i^{(h)} = 0$ otherwise. Then we calculate the coverage $C_i = \sum_{h=1}^{100} I_i^{(h)} / 100$ and $W_i = \sum_{h=1}^{100} W_i^{(h)} / 100$.

The biserial correlation (i.e., Pearson correlation coefficient) ρ between $\mathbf{y}$ and $\boldsymbol{\pi}$ is calculated since the variable $\mathbf{y}$ is binary (e.g., see Cox, 1974). The summaries based on AB, RMSE, coverage and width of the 95% HPD intervals for three correlation cases under the homogeneity assumption are shown in Table 4.9. Under this assumption, the performances of the HoS model is better than the HeS model in some domains, but the difference is very small. Both of them are better than the IS model in the sense of the closeness to the true P_i. In particular, we can see that as the correlation between $\mathbf{y}$ and $\boldsymbol{\pi}$ increases, there is greater discrepancy between the IS model and the others.

Table 4.10 shows the summaries for three correlation cases under the heterogeneity assumption. From this table, we find that the HeS model is well behaved in the sense that it is closer to true P_i than the other models when the effect of the selection bias is moderate to strong. It has smaller bias, smaller mean square error and better coverage.

As the biserial correlation increases there are increased disparities between the IS model and the HeS model. We present some plots of the posterior distributions of the finite population proportions for low, medium, high correlation cases under the homogeneity assumption (Figures 4.3-4.5). It is worth noting that there are large differences of the distributions of the finite population means as ρ increases. There are no differences in the posterior densities of the HoS model and the HeS model under the homogeneity assumption. The posterior distributions under the HeS model departs from the right of the distribution under the IS model, with little changes in their spreads for all domains.

Similarly, we present some plots of the posterior distributions of the finite population proportions for low, medium, high correlation cases under the heterogeneity assumption (Figures 4.6-4.8). The posterior densities under the HeS model are different from the others for all domains.

Also, we have selected two values of $\ell = 12, 24,$ the number of areas, to see changes by increasing ℓ. We compute AB, RMSE, coverage and width of 95% HPD intervals for each domain, then we average these values over all domains. We present the results under the homogeneity and the heterogeneity assumptions in respectively Tables 4.11 and 4.12. The results are consistent as ℓ increases.

The simulation study shows that as the biserial correlation between the binary responses and the selection probabilities increases, there is greater discrepancy among the IS model, the HoS moel and the HeS model. That is, as the correlation is stronger, the effect of the selection bias becomes larger. The HeS model is better than the other models when the each area has the different distribution for the sample selection probabilities under moderate to strong selection bias.

Table 4.9
Comparisons of estimates based on absolute bias (AB), root posterior mean squared error (RMSE), coverage (C) and width (W) of 95% HPD intervals for low (L), medium (M), high (H) correlation cases under the homogeneity assumption

area	ρ	AB			RMSE			W			C		
		IS	HoS	HeS	IS	HoS	HeS	IS	HoS	HeS	IS	HoS	HeS
1	L	0.061	0.044	0.051	0.087	0.076	0.087	0.241	0.248	0.281	0.90	0.98	0.98
	M	0.134	0.046	0.057	0.149	0.075	0.092	0.259	0.238	0.293	0.50	0.98	0.97
	H	0.188	0.047	0.050	0.198	0.075	0.089	0.268	0.239	0.296	0.16	0.99	0.99
2	L	0.050	0.053	0.051	0.081	0.087	0.093	0.257	0.277	0.315	0.98	0.98	0.98
	M	0.103	0.043	0.042	0.122	0.081	0.091	0.260	0.279	0.331	0.72	0.96	1.0
	H	0.166	0.045	0.051	0.177	0.084	0.100	0.254	0.288	0.350	0.24	0.99	1.0
3	L	0.051	0.065	0.063	0.083	0.098	0.102	0.262	0.286	0.320	0.97	0.90	0.98
	M	0.084	0.064	0.058	0.107	0.099	0.103	0.258	0.295	0.342	0.81	0.95	0.98
	H	0.132	0.056	0.059	0.146	0.094	0.108	0.250	0.304	0.361	0.45	0.97	0.97
4	L	0.058	0.084	0.085	0.091	0.114	0.120	0.269	0.297	0.333	0.91	0.88	0.89
	M	0.052	0.078	0.070	0.083	0.110	0.113	0.258	0.311	0.355	0.97	0.93	0.98
	H	0.112	0.066	0.072	0.128	0.105	0.118	0.243	0.325	0.372	0.57	0.97	0.98
5	L	0.039	0.037	0.041	0.073	0.074	0.083	0.250	0.263	0.296	0.99	0.99	1.0
	M	0.118	0.037	0.043	0.135	0.074	0.088	0.261	0.261	0.315	0.64	0.98	1.0
	H	0.193	0.050	0.055	0.202	0.083	0.100	0.261	0.269	0.336	0.14	0.99	0.99
6	L	0.052	0.040	0.046	0.081	0.075	0.085	0.245	0.256	0.289	0.96	0.96	0.99
	M	0.125	0.041	0.049	0.142	0.073	0.089	0.260	0.249	0.301	0.52	0.98	0.98
	H	0.186	0.048	0.056	0.197	0.078	0.096	0.263	0.253	0.316	0.18	0.99	0.98
7	L	0.041	0.053	0.058	0.076	0.088	0.098	0.261	0.281	0.316	0.97	0.95	0.96
	M	0.096	0.049	0.045	0.116	0.087	0.095	0.259	0.289	0.340	0.75	0.94	0.99
	H	0.145	0.046	0.050	0.158	0.086	0.102	0.254	0.293	0.356	0.36	0.96	0.98
8	L	0.044	0.039	0.045	0.076	0.075	0.085	0.249	0.261	0.295	0.99	0.99	1.0
	M	0.118	0.039	0.041	0.135	0.074	0.086	0.261	0.258	0.309	0.62	0.99	0.99
	H	0.184	0.049	0.058	0.195	0.081	0.101	0.262	0.262	0.329	0.24	0.98	0.97
9	L	0.047	0.043	0.046	0.077	0.078	0.088	0.253	0.268	0.301	0.98	0.98	0.99
	M	0.116	0.041	0.042	0.133	0.076	0.089	0.260	0.265	0.322	0.64	1.0	1.0
	H	0.184	0.050	0.059	0.194	0.084	0.103	0.258	0.274	0.338	0.21	0.98	0.97
10	L	0.062	0.046	0.059	0.088	0.078	0.095	0.246	0.257	0.296	0.94	0.98	0.97
	M	0.135	0.041	0.050	0.149	0.073	0.090	0.260	0.245	0.300	0.48	0.99	0.99
	H	0.191	0.055	0.058	0.202	0.082	0.097	0.264	0.247	0.310	0.17	0.99	1.0
11	L	0.053	0.070	0.074	0.085	0.103	0.112	0.265	0.292	0.327	0.96	0.92	0.95
	M	0.070	0.065	0.053	0.096	0.100	0.101	0.258	0.303	0.348	0.88	0.95	0.99
	H	0.131	0.058	0.058	0.145	0.098	0.109	0.245	0.313	0.368	0.46	0.94	0.97
12	L	0.047	0.043	0.043	0.077	0.077	0.086	0.250	0.264	0.306	0.97	0.99	0.99
	M	0.122	0.047	0.048	0.139	0.080	0.091	0.260	0.263	0.317	0.56	0.97	0.99
	H	0.173	0.044	0.050	0.184	0.078	0.095	0.261	0.266	0.333	0.27	0.99	0.99

Table 4.10
Comparisons of estimates based on absolute bias (AB), root posterior mean squared error (RMSE), coverage (C) and width (W) of 95% HPD intervals for low (L), medium (M), high (H) correlation cases under the heterogeneity assumption

area	ρ	AB			RMSE			W			C		
		IS	HoS	HeS	IS	HoS	HeS	IS	HoS	HeS	IS	HoS	HeS
1	L	0.050	0.052	0.047	0.078	0.085	0.084	0.215	0.235	0.251	0.91	0.92	0.94
	M	0.106	0.101	0.041	0.125	0.123	0.084	0.229	0.248	0.268	0.55	0.67	0.99
	H	0.163	0.145	0.056	0.176	0.162	0.109	0.242	0.261	0.343	0.16	0.39	0.99
2	L	0.045	0.047	0.050	0.079	0.085	0.094	0.231	0.252	0.287	0.95	0.97	0.99
	M	0.087	0.092	0.049	0.109	0.116	0.105	0.231	0.250	0.345	0.75	0.78	1.0
	H	0.161	0.174	0.090	0.173	0.187	0.145	0.221	0.235	0.382	0.26	0.23	0.89
3	L	0.056	0.058	0.068	0.088	0.094	0.108	0.235	0.259	0.287	0.91	0.92	0.91
	M	0.085	0.080	0.062	0.111	0.111	0.114	0.239	0.264	0.345	0.77	0.86	0.97
	H	0.151	0.133	0.060	0.165	0.152	0.129	0.233	0.261	0.405	0.32	0.55	0.97
4	L	0.056	0.059	0.056	0.088	0.095	0.099	0.240	0.263	0.290	0.92	0.93	0.96
	M	0.117	0.123	0.053	0.135	0.144	0.108	0.242	0.265	0.341	0.57	0.59	0.99
	H	0.175	0.194	0.081	0.188	0.207	0.140	0.245	0.267	0.404	0.23	0.20	0.96
5	L	0.047	0.049	0.047	0.078	0.084	0.089	0.224	0.245	0.273	0.98	0.97	0.99
	M	0.105	0.098	0.052	0.126	0.124	0.102	0.239	0.263	0.313	0.60	0.71	0.98
	H	0.164	0.144	0.071	0.178	0.163	0.130	0.246	0.271	0.399	0.28	0.46	1.0
6	L	0.045	0.051	0.059	0.076	0.084	0.097	0.218	0.238	0.273	0.95	0.95	0.95
	M	0.060	0.063	0.058	0.086	0.091	0.109	0.211	0.227	0.331	0.87	0.90	0.97
	H	0.103	0.113	0.055	0.118	0.128	0.112	0.195	0.203	0.353	0.51	0.53	0.99
7	L	0.043	0.046	0.056	0.079	0.085	0.098	0.234	0.257	0.284	0.97	0.98	0.97
	M	0.093	0.086	0.051	0.117	0.116	0.107	0.241	0.265	0.337	0.73	0.81	0.98
	H	0.152	0.134	0.065	0.168	0.155	0.129	0.240	0.266	0.400	0.31	0.53	0.98
8	L	0.045	0.048	0.060	0.077	0.083	0.098	0.223	0.244	0.277	0.96	0.99	0.96
	M	0.066	0.069	0.053	0.092	0.097	0.109	0.217	0.235	0.338	0.87	0.90	0.98
	H	0.103	0.115	0.064	0.121	0.132	0.123	0.207	0.217	0.375	0.56	0.48	0.99
9	L	0.048	0.050	0.045	0.080	0.086	0.088	0.227	0.249	0.275	0.98	0.97	0.99
	M	0.103	0.096	0.052	0.125	0.123	0.101	0.241	0.264	0.312	0.63	0.70	0.99
	H	0.166	0.146	0.057	0.180	0.166	0.120	0.245	0.272	0.382	0.24	0.42	0.99
10	L	0.046	0.047	0.057	0.076	0.080	0.095	0.214	0.235	0.272	0.92	0.96	0.95
	M	0.051	0.053	0.062	0.078	0.082	0.110	0.207	0.222	0.327	0.93	0.94	0.97
	H	0.093	0.102	0.054	0.108	0.117	0.110	0.191	0.198	0.348	0.61	0.60	1.0
11	L	0.057	0.058	0.071	0.088	0.094	0.110	0.237	0.260	0.293	0.91	0.95	0.92
	M	0.075	0.068	0.059	0.102	0.102	0.112	0.238	0.260	0.344	0.79	0.86	0.99
	H	0.136	0.120	0.067	0.152	0.142	0.128	0.229	0.257	0.398	0.43	0.58	0.99
12	L	0.044	0.046	0.051	0.076	0.082	0.092	0.222	0.243	0.277	0.98	0.96	1.0
	M	0.072	0.076	0.050	0.097	0.103	0.105	0.219	0.236	0.337	0.79	0.82	1.0
	H	0.115	0.127	0.064	0.131	0.143	0.123	0.211	0.222	0.369	0.51	0.45	0.98

Table 4.11
Comparison of the three models using absolute bias (AB), root posterior mean squared error (RMSE), coverage (C) and width (W) of 95% HPD intervals under the homogeneity assumption

ρ	ℓ	Model	AB	RMSE	W	C
Low	12	IS	$0.0503_{0.0075}$	$0.0812_{0.0056}$	$0.2539_{0.0086}$	$0.9600_{0.0292}$
		HoS	$0.0514_{0.0145}$	$0.0853_{0.0131}$	$0.2709_{0.0155}$	$0.9583_{0.0381}$
		HeS	$0.0551_{0.0133}$	$0.0943_{0.0117}$	$0.3062_{0.0161}$	$0.9733_{0.0303}$
	24	IS	$0.0486_{0.0040}$	$0.0809_{0.0029}$	$0.2594_{0.0123}$	$0.9721_{0.0177}$
		HoS	$0.0509_{0.0101}$	$0.0861_{0.0103}$	$0.2778_{0.0215}$	$0.9713_{0.0201}$
		HeS	$0.0568_{0.0105}$	$0.0975_{0.0101}$	$0.3166_{0.0210}$	$0.9792_{0.0177}$
Medium	12	IS	$0.1060_{0.0261}$	$0.1254_{0.0212}$	$0.2594_{0.0011}$	$0.6742_{0.1556}$
		HoS	$0.0491_{0.0127}$	$0.0835_{0.0126}$	$0.2714_{0.0237}$	$0.9683_{0.0221}$
		HeS	$0.0498_{0.0086}$	$0.0939_{0.0078}$	$0.3227_{0.0203}$	$0.9883_{0.0094}$
	24	IS	$0.1010_{0.0232}$	$0.1212_{0.0198}$	$0.2592_{0.0104}$	$0.7025_{0.1267}$
		HoS	$0.0502_{0.0092}$	$0.0861_{0.0110}$	$0.2812_{0.0295}$	$0.9775_{0.0159}$
		HeS	$0.0521_{0.0080}$	$0.0964_{0.0093}$	$0.3279_{0.0281}$	$0.9904_{0.0127}$
High	12	IS	$0.1653_{0.0281}$	$0.1771_{0.0263}$	$0.2568_{0.0078}$	$0.2875_{0.1399}$
		HoS	$0.0510_{0.0064}$	$0.0857_{0.0088}$	$0.2776_{0.0270}$	$0.9783_{0.0159}$
		HeS	$0.0562_{0.0062}$	$0.1014_{0.0075}$	$0.3389_{0.0237}$	$0.9825_{0.0114}$
	24	IS	$0.1582_{0.0316}$	$0.1704_{0.0300}$	$0.2540_{0.0170}$	$0.3242_{0.1618}$
		HoS	$0.0522_{0.0046}$	$0.0886_{0.0090}$	$0.2882_{0.0335}$	$0.9788_{0.0099}$
		HeS	$0.0570_{0.0039}$	$0.1025_{0.0069}$	$0.3412_{0.0302}$	$0.9875_{0.0126}$

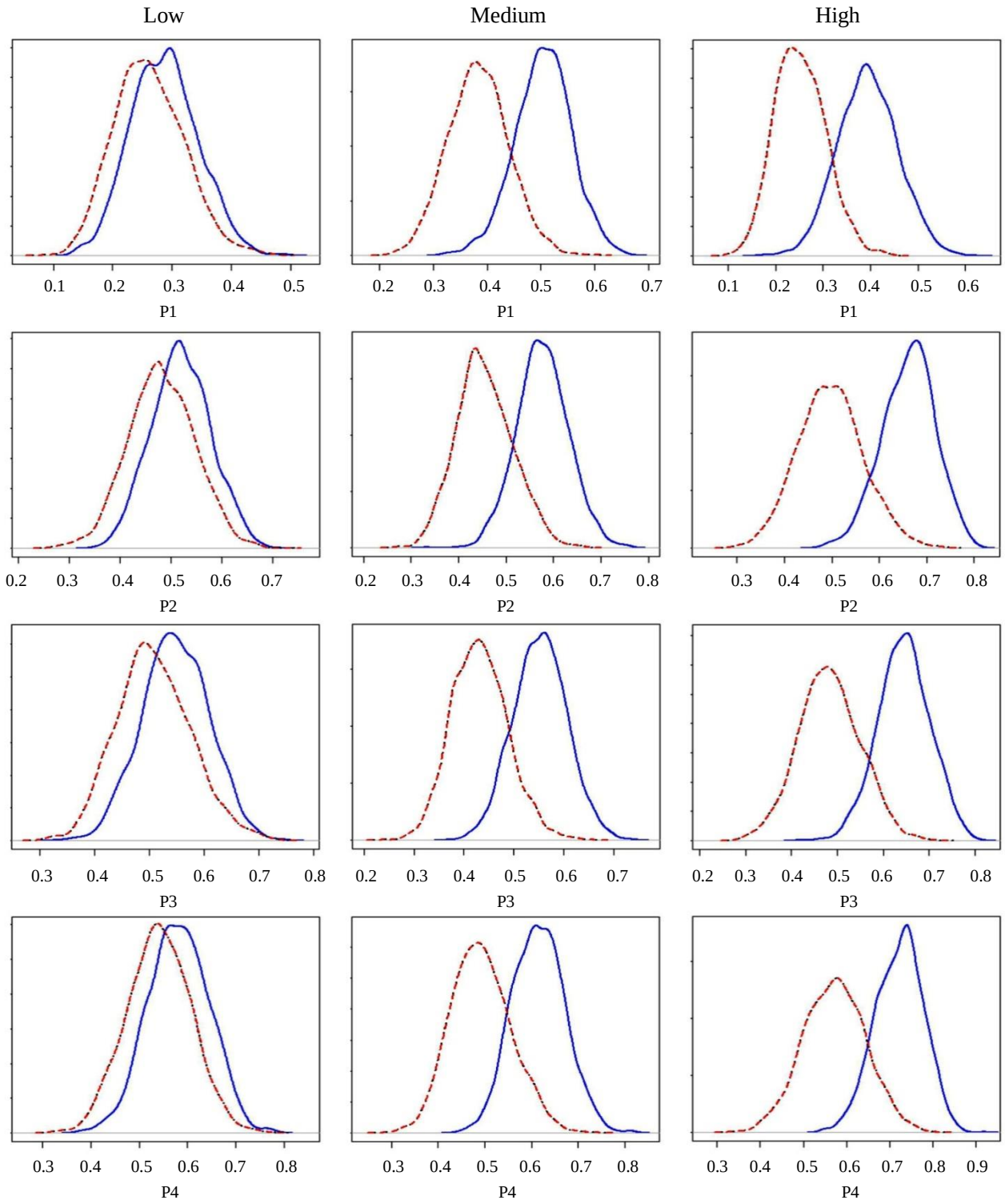
NOTE: The notation a_b means a is an average of areas and b is the standard error.

Table 4.12
Comparison of the three models using absolute bias (AB), root posterior mean squared error (RMSE), coverage (C) and width (W) of 95% HPD intervals under the heterogeneity assumption

ρ	ℓ	Model	AB	RMSE	W	C
Low	12	IS	$0.0501_{0.0050}$	$0.0842_{0.0051}$	$0.2703_{0.0109}$	$0.9733_{0.0156}$
		HoS	$0.0509_{0.0049}$	$0.0862_{0.0049}$	$0.2800_{0.0112}$	$0.9758_{0.0168}$
		HeS	$0.0556_{0.0083}$	$0.0959_{0.0075}$	$0.3132_{0.0127}$	$0.9767_{0.0227}$
	24	IS	$0.0457_{0.0054}$	$0.0768_{0.0054}$	$0.2478_{0.0135}$	$0.9675_{0.0205}$
		HoS	$0.0470_{0.0055}$	$0.0818_{0.0053}$	$0.2703_{0.0146}$	$0.9779_{0.0169}$
		HeS	$0.0537_{0.0093}$	$0.0922_{0.0082}$	$0.3018_{0.0167}$	$0.9792_{0.0193}$
Medium	12	IS	$0.0850_{0.0206}$	$0.1085_{0.0179}$	$0.2586_{0.0147}$	$0.8033_{0.1026}$
		HoS	$0.0837_{0.0196}$	$0.1109_{0.0170}$	$0.2816_{0.0187}$	$0.8517_{0.0720}$
		HeS	$0.0534_{0.0060}$	$0.1054_{0.0077}$	$0.3672_{0.0248}$	$0.9917_{0.0111}$
	24	IS	$0.0824_{0.0218}$	$0.1056_{0.0190}$	$0.2531_{0.0179}$	$0.7992_{0.1040}$
		HoS	$0.0810_{0.0205}$	$0.1074_{0.0179}$	$0.2736_{0.0218}$	$0.8546_{0.0825}$
		HeS	$0.0533_{0.0077}$	$0.1046_{0.0090}$	$0.3637_{0.0289}$	$0.9954_{0.0093}$
High	12	IS	$0.1401_{0.0289}$	$0.1547_{0.0278}$	$0.2537_{0.0232}$	$0.4417_{0.1333}$
		HoS	$0.1372_{0.0263}$	$0.1543_{0.0253}$	$0.2741_{0.0319}$	$0.5375_{0.1231}$
		HeS	$0.0653_{0.0110}$	$0.1249_{0.0113}$	$0.4215_{0.0257}$	$0.9833_{0.0287}$
	24	IS	$0.1336_{0.0302}$	$0.1481_{0.0289}$	$0.2487_{0.0263}$	$0.4746_{0.1316}$
		HoS	$0.1302_{0.0270}$	$0.1470_{0.0261}$	$0.2663_{0.0348}$	$0.5567_{0.1396}$
		HeS	$0.0601_{0.0090}$	$0.1203_{0.0108}$	$0.4141_{0.0360}$	$0.9900_{0.0153}$

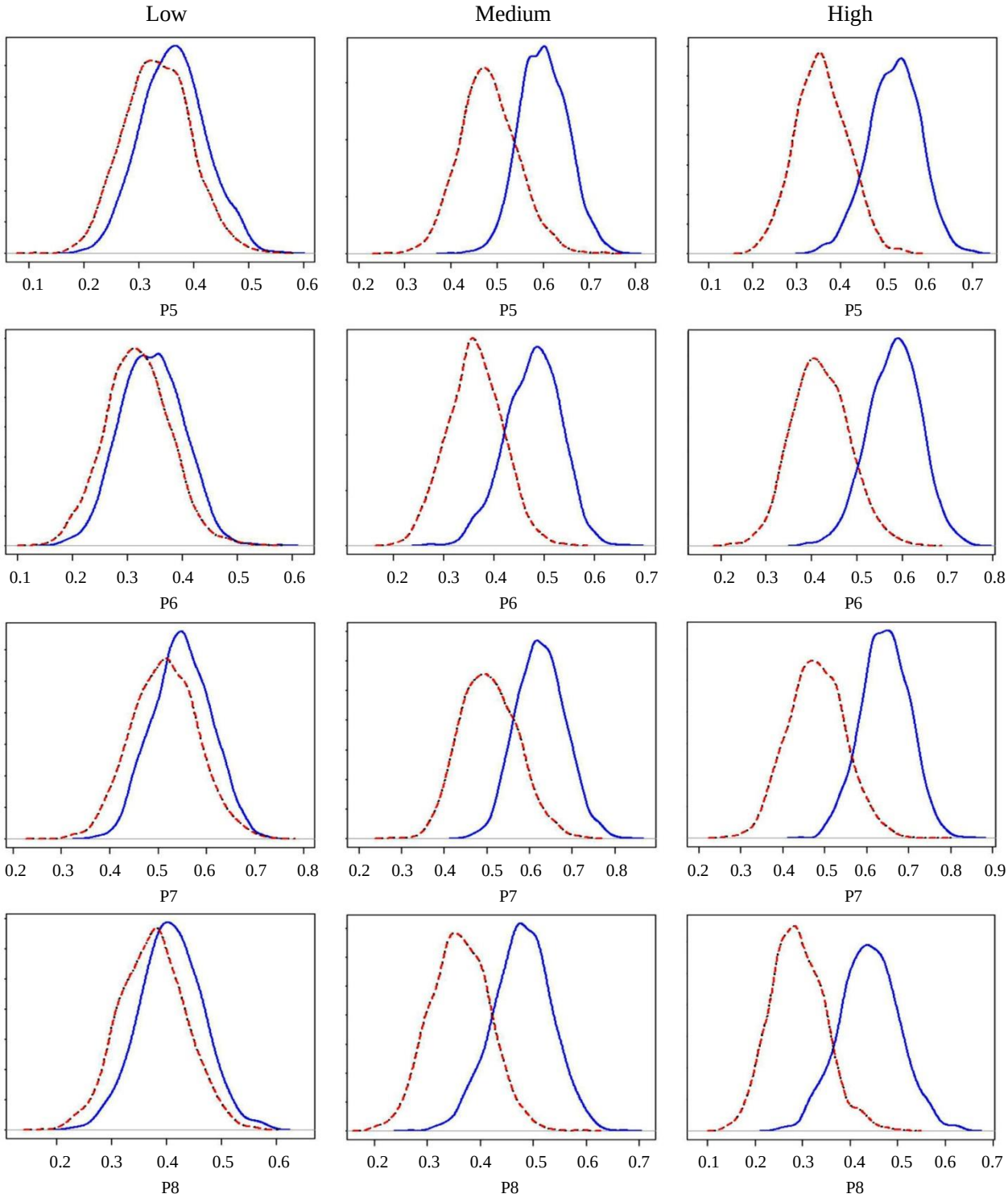
NOTE: The notation a_b means a is an average of areas and b is the standard error.

Figure 4.3 The posterior densities of the finite population proportions for low, medium, high correlation cases using a simulated data under the homogeneity assumption.



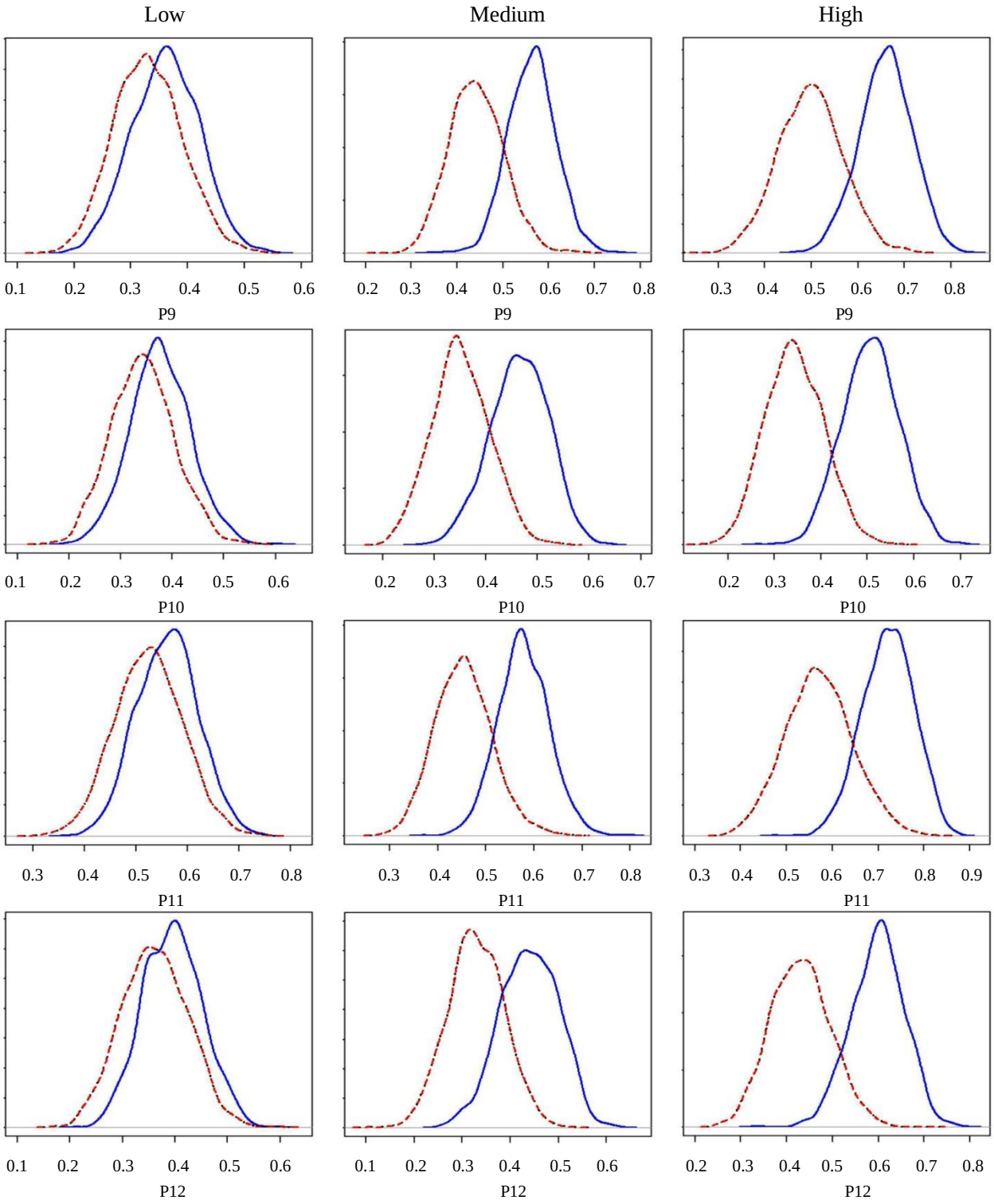
NOTE: Here the solid line represents the densities of P under the IS model, the dotted line represents that under the HoS model, and the dashed line represents that under the HeS model. (Dashed line and dotted line coincide.)

Figure 4.4 The posterior densities of the finite population proportions for low, medium, high correlation cases using a simulated data under the homogeneity assumption (continued).



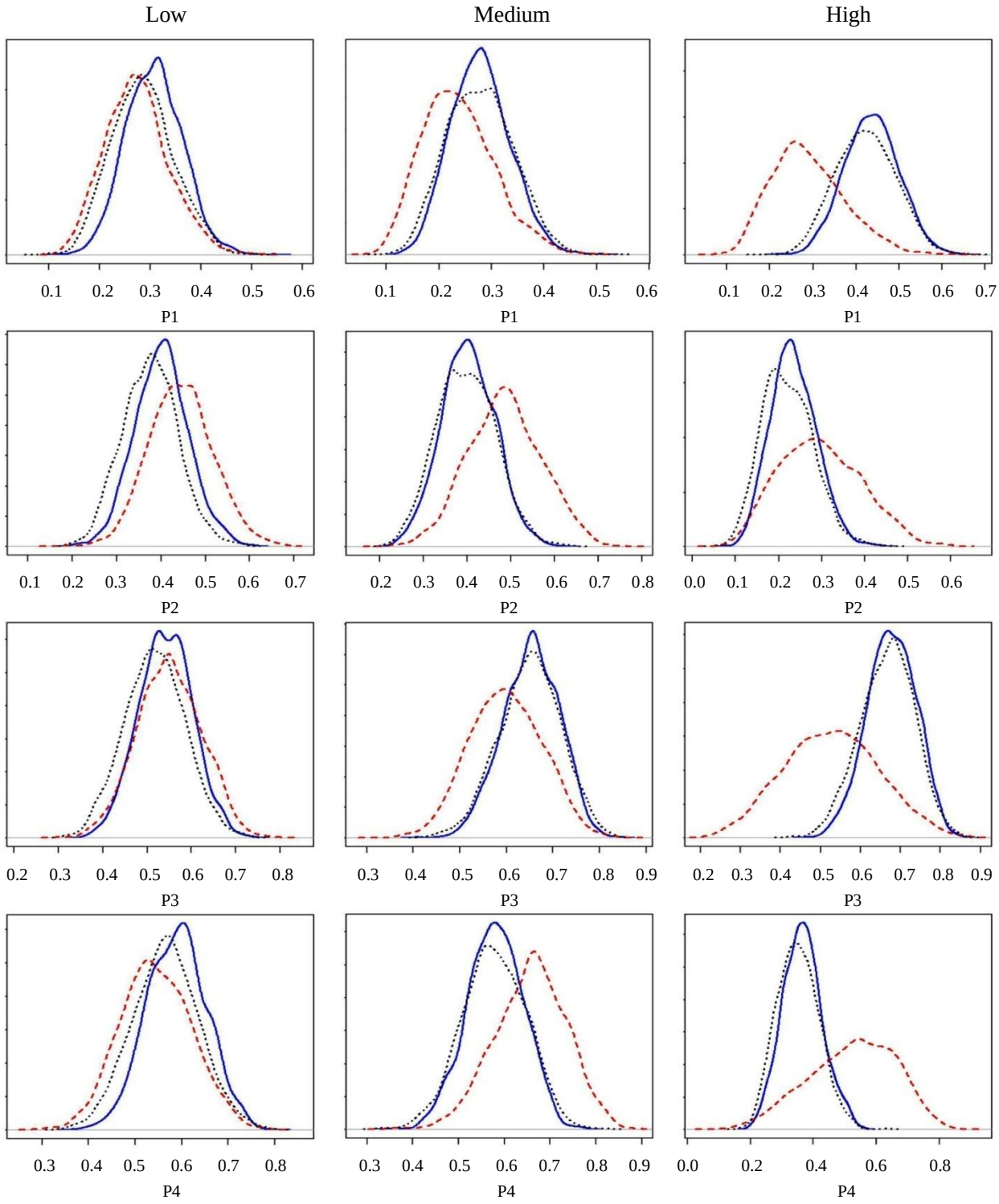
NOTE: Here the solid line represents the densities of P under the IS model, the dotted line represents that under the HoS model, and the dashed line represents that under the HeS model. (Dashed line and dotted line coincide.)

Figure 4.5 The posterior densities of the finite population proportions for low, medium, high correlation cases using a simulated data under the homogeneity assumption (continued).



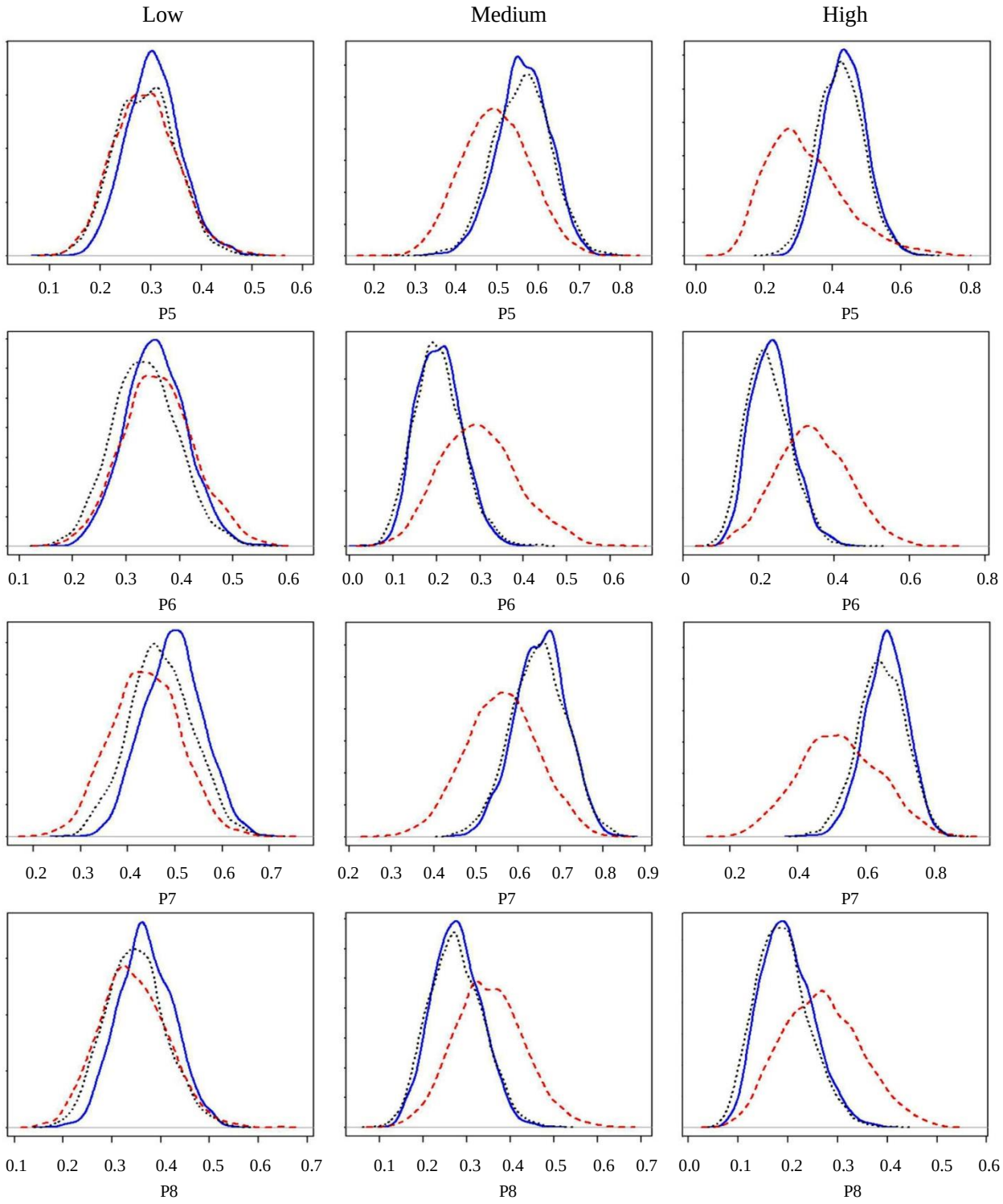
NOTE: Here the solid line represents the densities of P under the IS model, the dotted line represents that under the HoS model, and the dashed line represents that under the HeS model. (Dashed line and dotted line coincide.)

Figure 4.6 The posterior densities of the finite population proportions for low, medium, high correlation cases using a simulated data under the heterogeneity assumption.



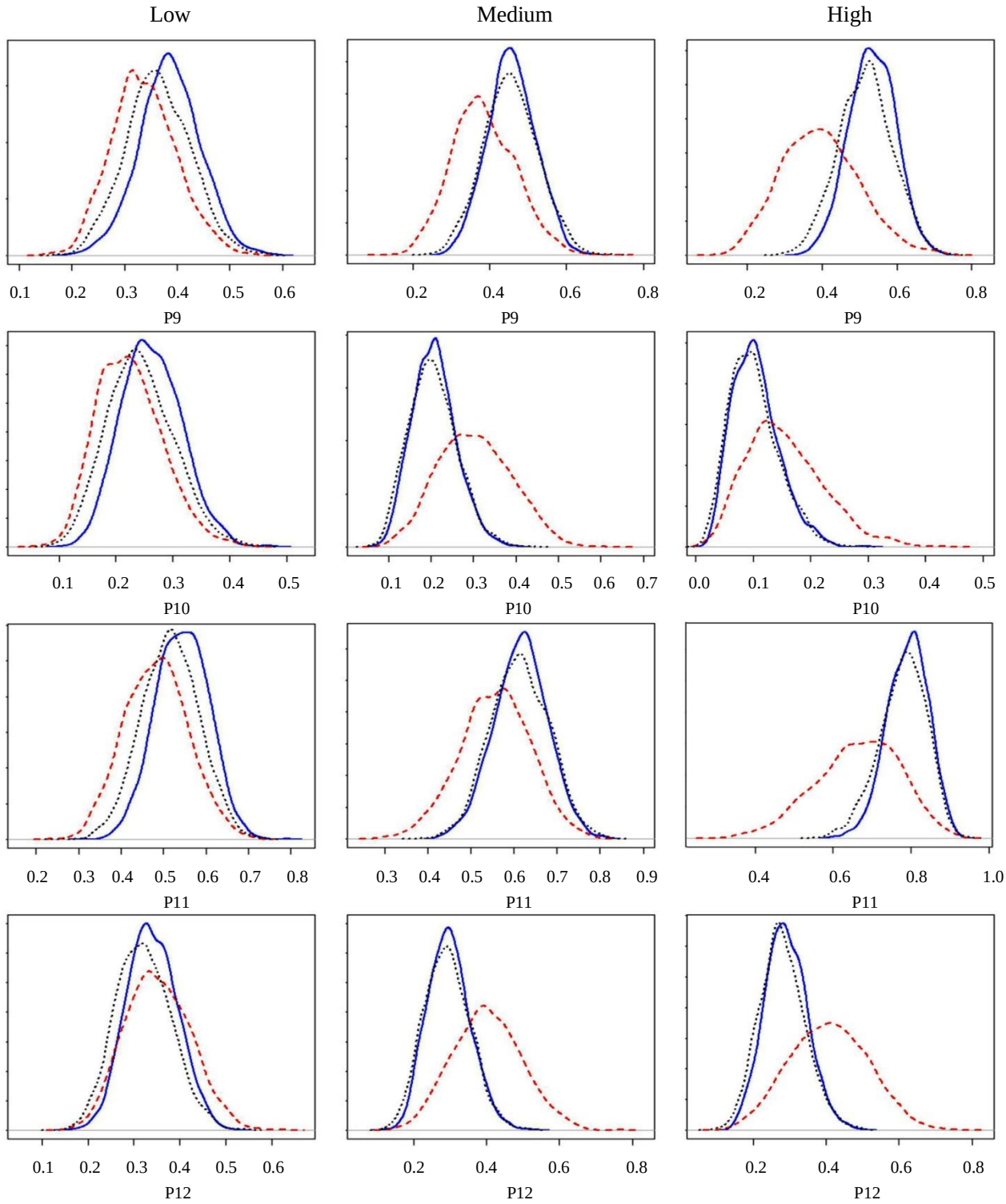
NOTE: Here the solid line represents the densities of P under the IS model, the dotted line represents that under the HoS model, and the dashed line represents that under the HeS model.

Figure 4.7 The posterior densities of the finite population proportions for low, medium, high correlation cases using a simulated data under the heterogeneity assumption (continued).



NOTE: Here the solid line represents the densities of P under the IS model, the dotted line represents that under the HoS model, and the dashed line represents that under the HeS model.

Figure 4.8 The posterior densities of the finite population proportions for low, medium, high correlation cases using a simulated data under the heterogeneity assumption (continued).



NOTE: Here the solid line represents the densities of P under the IS model, the dotted line represents that under the HoS model, and the dashed line represents that under the HeS model.

5. Concluding remarks

We have extended the homogeneous nonignorable selection model of CNK to accommodate selection probabilities that have different distributions in different areas. This is an improvement to handle the relationship between the binary variables and the selection probabilities. As there are numerous additional parameters, the computation has become a lot more difficult and we have used the Metropolis-Hastings sampler to overcome this difficulty.

We have used an example on severe activity limitation in the National Health Interview Survey in which the small areas are formed by crossing education (pre-college, college, post-college), sex (male, female) and race (white, nonwhite). The heterogeneous nonignorable selection model appears to perform better than the homogeneous selection model and the baseline ignorable selection model.

We have used a simulation study to assess the performance of our heterogeneous nonignorable selection model. We have drawn data from the homogeneous nonignorable selection model and fitted the ignorable selection model and the heterogeneous nonignorable selection model. We found little difference between the two nonignorable selection models but substantial difference from the ignorable selection model. However, when we have drawn data from the heterogeneous nonignorable selection model and fitted the ignorable selection model and the homogeneous nonignorable selection model, we found that the heterogeneous nonignorable selection model is a lot better especially when there is medium to strong selection bias using bias, mean square error and coverage. This is evident in Tables 4.9-4.12 and Figures 4.3-4.8.

Within the framework of our heterogeneous nonignorable selection model, we can think about several additional problems. First, we can accommodate polychotomous data that are numerous in survey problems. Second, although a bit different from our approach, we can accommodate covariates (e.g., age, race, sex in our application on activity limitation). Third, as there are nonrespondents in numerous surveys, we can attempt to accommodate nonignorable nonresponse and selection simultaneously (e.g., Nandram and Choi, 2010). This can be done within our framework. Fourth, as in a two-fold model, we can accommodate clustering or stratification within each area (e.g., Nandram, 2016; Lee, Nandram and Kim, 2017). Our work is potentially useful to solve problems in nonprobability samples as well (e.g., Elliot and Valliant, 2017).

Acknowledgements

This research was supported by a grant from the Simons Foundation (#353953, Balgobin Nandram).

References

Chambers, R., Dorfman, A. and Wang, S. (1998). Limited information likelihood analysis of survey data. *Journal of the Royal Statistical Society, Series B*, 60, 397-411.

Chen, Q., Elliott, M.R. and Little, R.J.A. (2010). Bayesian penalized spline model-based inference for finite population proportion in unequal probability sampling. *Survey Methodology*, 36, 1, 23-34. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2010001/article/11250-eng.pdf.

Choi, S., Nandram, B. and Kim, D. (2017). A hierarchical Bayesian model for binary data incorporating selection bias. *Communications in Statistics - Simulation and Computation*, 46, 6, 4767-4782.

Cox, N.R. (1974). Estimation of the correlation between a continuous and a discrete variable. *Biometrics*, 30, 171-178.

Elliot, M.R., and Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32, 249-264.

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

Lee, D., Nandram, B. and Kim, D. (2017). Bayesian predictive inference of a proportion under a two-fold small area model with heterogeneous correlations. *Survey Methodology*, 43, 1, 69-92. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2017001/article/14822-eng.pdf.

Ma, J., Sedransk, J., Nandram, B. and Chen, L. (2018). Bayesian predictive inference for finite population quantities under informative sampling. *Statistics and Application*, 16, 207-226.

Malec, D., Davis, W. and Cao, X. (1999). Model-based small area estimates of overweight prevalence using sample selection adjustment. *Statistics in Medicine*, 18, 3189-3200.

Nandram, B. (2007). Bayesian predictive inference under informative sampling via surrogate samples. *Bayesian Statistics and its Applications*, (Eds. S.K. Upadhyay, U. Singh and D. Dey), Anamaya, New Delhi, Chapter 25, 356-374.

Nandram, B. (2016). Bayesian predictive inference of a proportion under a two-fold small area model. *Journal of Official Statistics*, 32, 1, 187-208.

Nandram, B., and Choi, J.W. (2010). A Bayesian analysis of body mass index data from small domains under nonignorable nonresponse and selection. *Journal of the American Statistical Association*, 105, 120-135.

Nandram, B., Bai, Y. and Choi, J.W. (2011). Hierarchical Bayesian models for assessing possible changes in prevalence of activity limitation. *Advances and Applications in Statistical Sciences*, 6, 285-311.

Nandram, B., Bhatta, D., Bhadra, D. and Shen G. (2013). Bayesian predictive inference of a finite population proportion under selection bias. *Statistical Methodology*, 11, 1-21.

Nandram, B., Choi, J.W., Shen, G. and Burgos, C. (2006). Bayesian predictive inference under informative sampling and transformation. *Applied Stochastic Models in Business and Industry*, 22, 559-572.

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

Pfeffermann, D. (1988). The effect of sampling design and response mechanism on multivariate regression-based predictors. *Journal of the American Statistical Association*, 83, 824-833.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61, 317-337.

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

Pfeffermann, D., Krieger, A.M. and Rinott, Y. (1998). Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica*, 8, 1087-1114.

Si, Y., Pillai, N.S. and Gelman, A. (2015). Bayesian nonparametric weighted sampling inference. *Bayesian Analysis*, 10, 3, 605-625.

Sverchkov, M., and Pfeffermann, D. (2004). Prediction of finite population totals based on the sample distribution. *Survey Methodology*, 30, 1, 79-92. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2004001/article/6996-eng.pdf.

Zangeneh, S.Z., and Little, R.J.A. (2015). Bayesian inference for the finite population total from a heteroscedastic probability proportional to size sample. *Journal of Survey Statistics and Methodology*, 3, 162-192.

Survey Methodology, June 2021
Vol. 47, No. 1, pp. 123-149
Statistics Canada, Catalogue No. 12-001-X

Small area benchmarked estimation under the basic unit level model when the sampling rates are non-negligible

Marius Stefan and Michael A. Hidiroglou[1]

Abstract

We consider the estimation of a small area mean under the basic unit-level model. The sum of the resulting model-dependent estimators may not add up to estimates obtained with a direct survey estimator that is deemed to be accurate for the union of these small areas. Benchmarking forces the model-based estimators to agree with the direct estimator at the aggregated area level. The generalized regression estimator is the direct estimator that we benchmark to. In this paper we compare small area benchmarked estimators based on four procedures. The first procedure produces benchmarked estimators by ratio adjustment. The second procedure is based on the empirical best linear unbiased estimator obtained under the unit-level model augmented with a suitable variable that ensures benchmarking. The third procedure uses pseudo-empirical estimators constructed with suitably chosen sampling weights so that, when aggregated, they agree with the reliable direct estimator for the larger area. The fourth procedure produces benchmarked estimators that are the result of a minimization problem subject to the constraint given by the benchmark condition. These benchmark procedures are applied to the small area estimators when the sampling rates are non-negligible. The resulting benchmarked estimators are compared in terms of relative bias and mean squared error using both a design-based simulation study as well as an example with real survey data.

Key Words: Small area; Benchmarking; Empirical estimator; Pseudo-empirical estimator; Constrained estimator.

1. Introduction

Small area estimation (SAE) has grown in importance in recent years due to the demand for reliable small area statistics. Direct estimators are used to estimate parameters of interest when the sample size is reasonably large. However, they have large standard errors and coefficients of variation when it comes to applying them to small areas, as the realized sample size will be quite small. It is therefore necessary to use models that borrow strength from other related areas or from past surveys to have stable estimators for these small areas. Model-based estimates typically show a substantial improvement over direct estimates in terms of mean squared error (MSE).

The available theory for small area estimation is based on either area-level or unit-level models, depending on the level of available auxiliary information. Unit-level based methods use the data of the individual units as auxiliary information, whereas area-level based methods use aggregates or means of the data of the units within the small areas. Fay and Herriot (1979), denoted hereafter as the FH model, is the most used area-level model in small area estimation. The one-fold nested error regression model proposed in Battese, Harter and Fuller (1988), also known as the basic unit-level model, is frequently used when unit-level information is available. We denote this model as the BHF model. Both are special cases of a general linear mixed model in SAE (see Rao and Molina, 2015 for an excellent account of the small area estimation).

Small area means or totals are the most frequent linear parameters estimated in SAE. In these cases, the most popular small area method is the use of linear mixed models to derive the best linear unbiased

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

predictors (BLUP) for the small area mean or total. BLUP estimators minimize the MSE among the class of linear unbiased estimators. Alternatively, it can be shown that the BLUP estimator can be obtained by solving mixed model equations with unknowns given by the fixed and random parameters of the model. The mixed model equations result from the maximization of the joint density of the data and the vector of random small area effects. A BLUP estimator depends on the variances (and covariances) of random effects which can be estimated by the Henderson method of fitting constants (FC), the maximum likelihood (ML) or restricted maximum likelihood (REML). Using these estimated components in the BLUP estimator leads to a two-stage estimator referred to as the empirical best linear unbiased predictor (EBLUP).

A potential difficulty with EBLUP estimators is that when they are aggregated over all the small areas, they may not agree with the overall estimate for a larger area obtained via direct estimation. Statistical agencies favor an overall agreement between the sum of the model-based small area estimates and the direct estimate at a higher level that corresponds to the union of the small areas. Benchmarking is a method of modifying the model-based estimates to agree with the direct estimator for the larger area.

Existing benchmarking methods are either frequentist or Bayesian. In this paper, we focus on the frequentist approach to benchmarking (for Bayesian benchmarking procedures, see You, Rao and Dick, 2004; Datta, Ghosh, Steorts and Maples, 2011 and Nandram and Sayit, 2011). The frequentist methods can be applied to obtain benchmark small area estimates for both the area-level and unit-level models.

We briefly summarize the existing literature for both types of models. We first describe the procedures developed to benchmark area-level based estimates. Pfeffermann and Barnard (1991) obtained a constrained benchmarked estimator by maximizing the joint density of the data and the vector of random small area effects given the benchmark restriction. Their benchmark estimator was constructed with modified estimates of fixed and small area effects that are solutions to the constrained maximization problem. Wang, Fuller and Qu (2008) developed a benchmarked EBLUP for the FH area-level model, by minimizing a loss function subject to the constraint given by the benchmark condition. They obtained a second benchmarked estimator by adding a suitable auxiliary variable to the FH model without imposing a constraint. They showed that the EBLUP estimator based on the augmented FH model is self-benchmarked: the estimator satisfied the benchmark condition without further adjustments. Bell, Datta and Ghosh (2013) generalized the result in Wang et al. (2008) to the case of multiple benchmark constraints by considering a more general loss function. You, Rao and Hidiroglou (2013) obtained another self-benchmarked estimator under the FH model by replacing the regression vector used in the EBLUP estimator with an alternative estimator that depends on the benchmarking weights.

We now turn to procedures that benchmark unit-level model-based estimates. The objective is to obtain small area estimators that benchmark to a direct estimator at a given level of aggregation of the small areas. The direct estimators that are mostly used by Statistical agencies are the Generalized Regression Estimator (GREG) in Särndal, Swensson and Wretman (1989) or more generally the calibration estimator based on procedures in Deville and Särndal (1992). You and Rao (2002) developed a pseudo-EBLUP predictor (YR predictor) that incorporates survey weights. A property of this estimator is that it is

self-benchmarked, that is, the sum of the small area estimates adds up to an estimator that has the same form as the GREG. However, it is not a direct estimator because the estimated regression vector that is part of this estimator reflects the error structure of the nested error model. Assuming that the sampling rates are negligible, Stefan and Hidiroglou (2020) proposed several procedures to ensure that the EBLUP and pseudo-EBLUP estimators would benchmark to the GREG estimator, given that both the model and the GREG estimator used the same vector of auxiliary variables. Ugarte, Militino and Goicoa (2009) developed a restricted EBLUP estimator for a small area total that satisfies the benchmarking property to a synthetic estimator.

The objective of this paper is to compare several benchmarked estimators of a small area mean for the basic unit level model when the sampling rates are non-negligible. We compare six benchmarked estimators: two benchmarked estimators based on the procedures proposed by Stefan and Hidiroglou (2020), two restricted estimators based on the procedure proposed by Ugarte et al. (2009) and two ratio estimators obtained by multiplying each small area EBLUP and YR estimators by a common adjustment factor. The paper is organized as follows. Section 2 presents a summary of EBLUP and pseudo-EBLUP estimators under the basic unit-level model. Section 3 describes the six benchmarked estimators. The first two estimators are based on simple ratio adjustments. Then, we show how the two benchmarking procedures proposed by Stefan and Hidiroglou (2020) in the case of negligible sampling rates can be adapted to produce benchmarked small area mean estimators when the sampling rates are non-negligible. Finally, we describe the restricted EBLUP estimator of Ugarte et al. (2009) and propose a pseudo restricted estimator which is a variant of the restricted EBLUP that incorporates survey weights. We also propose a re-parameterized restricted maximum likelihood (reREML) method for estimating the variance components. This method of estimation is useful when computing restricted EBLUP small area mean estimators as it results in strictly positive variance components estimates. Section 4 presents the results of a Monte Carlo simulation based on generated data sets, whereas Section 5 reports the results of a simulation study based on a real data set. Finally, Section 6 gives some concluding remarks.

2. EBLUP and pseudo-EBLUP estimation

Consider the one-fold nested error regression model

$$y_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + v_i + e_{ij},\ i = 1, \ldots, m;\ j = 1, \ldots, N_i, \tag{2.1}$$

where y_{ij} is the variable of interest for the j^{th} population unit in the i^{th} small area, $\mathbf{x}_{ij} = \left(x_{ij1}, \ldots, x_{ijp}\right)^T$ is a vector of auxiliary variables with $x_{ij1} = 1$, $\boldsymbol{\beta} = \left(\beta_1, \ldots, \beta_p\right)^T$ is a $p \times 1$ vector of regression parameters and N_i is the number of population units in the i^{th} small area, U_i. The random small area effects v_i are assumed to be i.i.d. $N\left(0, \sigma_v^2\right)$, and independent of the unit errors e_{ij}, which are assumed i.i.d. $N\left(0, \sigma_e^2\right)$. We draw samples s_i of size n_i independently within each small area i, according to a specified sampling design with first-order inclusion probabilities denoted by π_{ij}, for $j = 1, \ldots, N_i$. The total sample size is n, where $n = \sum_{i=1}^{m} n_i$. The resulting basic design weights are given by $d_{ij} = 1/\pi_{ij}$.

We assume that the sample design is ignorable, and that selection bias is absent. This implies that model (2.1) also holds for the sample data:

$$y_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + v_i + e_{ij},\ i = 1, \ldots, m;\ j = 1, \ldots, n_i, \tag{2.2}$$

Model (2.2) is a special case of the general linear mixed model. Defining $\mathbf{y}_i = \left(y_{i1}, \ldots, y_{in_i}\right)^T$, $\mathbf{X}_i = \left(\mathbf{x}_{i1}^T, \ldots, \mathbf{x}_{in_i}^T\right)^T$, $\mathbf{v} = (v_1, \ldots, v_m)^T$ and $\mathbf{e}_i = \left(e_{i1}, \ldots, e_{in_i}\right)^T$, it follows that model (2.2) can be expressed in a matrix form by stacking the observations. The resulting equation is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \mathbf{e}, \tag{2.3}$$

where $\mathbf{y} = \text{col}_{1\le i\le m}(\mathbf{y}_i)$, $\mathbf{X} = \text{col}_{1\le i\le m}(\mathbf{X}_i)$, $\mathbf{Z} = \text{diag}_{1\le i\le m}\{\mathbf{1}_{n_i}\}$ and $\mathbf{e} = \text{col}_{1\le i\le m}(\mathbf{e}_i)$, with $\mathbf{1}_{n_i}$ a vector of dimension n_i composed of ones. We denote by $\mathbf{G}$ and $\mathbf{R}$ the variance matrices of the random vectors $\mathbf{v}$ and $\mathbf{e}$ respectively. Then $\mathbf{G} = \sigma_v^2\mathbf{I}_m$ and $\mathbf{R} = \sigma_e^2\mathbf{I}_n$. It follows that the variance matrix of vector $\mathbf{y}$, denoted as $\mathbf{V}$, is given by $\mathbf{V} = \mathbf{R} + \mathbf{Z}\mathbf{G}\mathbf{Z}^T$.

The parameters of interest are the small area means $\bar{Y}_i$, where $\bar{Y}_i = N_i^{-1}\sum_{j=1}^{N_i} y_{ij}$, $i = 1, \ldots, m$. If N_i is large, the sampling fraction $f_i = N_i^{-1}n_i$, of the i^{th} small area is negligible. This set-up corresponds to the case of an *infinite population* or negligible sampling rates. It follows that the small area means $\bar{Y}_i$ can be approximated by μ_i (see Rao and Molina, 2015, page 174), where $\mu_i = \bar{\mathbf{X}}_i^T\boldsymbol{\beta} + v_i$ and $\bar{\mathbf{X}}_i = \sum_{j=1}^{N_i}\mathbf{x}_{ij}/N_i$ is the vector of population means of the $\mathbf{x}_{ij}$'s for the i^{th} area. An estimator of μ_i is given by $\hat{\mu}_i = \bar{\mathbf{X}}_i^T\hat{\boldsymbol{\beta}} + \hat{v}_i$ (Rao and Molina, 2015, page 175), where $\hat{\boldsymbol{\beta}}$ and $\hat{v}_i$ are estimators of $\boldsymbol{\beta}$ and v_i respectively. If N_i is not large enough or if the sampling rates f_i are not negligible, parameters $\bar{Y}_i$ cannot be approximated by linear combinations of $\boldsymbol{\beta}$ and v_i. This corresponds to the case of a *finite population*. Let r_i be the set of the $N_i - n_i$ unobserved y-values in small area i. If we assume that we know the $\mathbf{x}_{ij}$'s for each individual in the population, an estimator $\hat{\bar{Y}}_i$ of $\bar{Y}_i$ is based on the observed values y_{ij}, $j \in s_i$, and predicted values $\hat{y}_{ij} = \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}} + \hat{v}_i$ for $j \in r_i$. That is, estimator $\hat{\bar{Y}}_i$ is given by

$$\hat{\bar{Y}}_i = \frac{1}{N_i}\left(\sum_{j\in s_i} y_{ij} + \sum_{j\in r_i}\hat{y}_{ij}\right). \tag{2.4}$$

Much of the SAE theory deals with the infinite population case, whereas the literature on the finite population case is more limited. In this paper we focus on finite population (or non-negligible sampling rates) case, thereby constructing estimators based on (2.4).

2.1 EBLUP estimation

We denote by $\tilde{\boldsymbol{\beta}}$ and $\tilde{\mathbf{v}} = (\tilde{v}_1, \ldots, \tilde{v}_m)^T$ the BLUP predictors of $\boldsymbol{\beta}$ and $\mathbf{v}$ respectively. These estimators are given by $\tilde{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$ and $\tilde{\mathbf{v}} = \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}\right)$. Under the normality assumption of $\mathbf{e}$ and $\mathbf{v}$, it can be shown that $\tilde{\boldsymbol{\beta}}$ and $\tilde{\mathbf{v}}$ can be obtained by maximizing the joint density of $\mathbf{y}$ and $\mathbf{v}$ with respect to $\boldsymbol{\beta}$ and $\mathbf{v}$. This is equivalent to minimizing the function

$$\phi = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{v})^T\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{v}) + \mathbf{v}^T\mathbf{G}^{-1}\mathbf{v}. \tag{2.5}$$

This leads to the following mixed model equations

$$\mathbf{A}\begin{pmatrix}\tilde{\boldsymbol{\beta}}\\ \tilde{\mathbf{v}}\end{pmatrix} = \mathbf{b}, \tag{2.6}$$

where

$$\mathbf{A} = \begin{pmatrix}\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z}\\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1}\end{pmatrix} \quad \text{and} \quad \mathbf{b} = \begin{pmatrix}\mathbf{X}^T\mathbf{R}^{-1}\mathbf{y}\\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{y}\end{pmatrix}. \tag{2.7}$$

(see Rao and Molina, 2015, page 99 for details). The variance components (σ_v^2, σ_e^2) in equations (2.6) and (2.7) are generally unknown. Three methods of estimation, FC, ML and REML, are commonly used in SAE to estimate the variance components (σ_v^2, σ_e^2). A well-known difficulty with these methods is that the estimate of σ_v^2 can take on negative values. This estimate is truncated to zero when this occurs, that is $\hat{\sigma}_v^2$ is set to 0. Empirical versions of $\mathbf{A}$ and $\mathbf{b}$, denoted as $\hat{\mathbf{A}}$ and $\hat{\mathbf{b}}$, are obtained if the unknown variance components (σ_v^2, σ_e^2) are replaced by estimators $(\hat{\sigma}_v^2, \hat{\sigma}_e^2)$. It follows from equation (2.6) that EBLUP estimators of model parameters $(\boldsymbol{\beta}, \mathbf{v})$, denoted as $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{v}} = (\hat{v}_1, \ldots, \hat{v}_m)^T$, are given by

$$\begin{pmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\mathbf{v}}\end{pmatrix} = \hat{\mathbf{A}}^{-1}\hat{\mathbf{b}}. \tag{2.8}$$

Using (2.8), it can be proved that $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{v}}$ are

$$\begin{pmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\mathbf{v}}\end{pmatrix} = \begin{pmatrix}\left(\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{y}\\ \hat{\mathbf{G}}\mathbf{Z}^T\hat{\mathbf{V}}^{-1}\left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)\end{pmatrix}, \tag{2.9}$$

where $\hat{\mathbf{G}} = \hat{\sigma}_v^2\mathbf{I}_m$ and $\hat{\mathbf{V}} = \hat{\sigma}_e^2\mathbf{I}_n + \hat{\sigma}_v^2\mathbf{Z}\mathbf{Z}^T$.

Remark 1. It is easier to invert matrices $\hat{\mathbf{G}} = \hat{\sigma}_v^2\mathbf{I}_m$ and $\hat{\mathbf{R}} = \hat{\sigma}_e^2\mathbf{I}_n$ than $\hat{\mathbf{V}}$. Consequently, it is simpler to use the mixed model equations (2.8) than equations (2.9) for computing $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{v}}$. However, when $\hat{\sigma}_v^2$ is equal to zero, equations (2.8) cannot be used because the $\hat{\mathbf{G}}^{-1}$ term in matrix $\hat{\mathbf{A}}$ does not exist. In such cases, $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{v}}$ can only be computed using (2.9).

Under model (2.2), it can be shown that $\hat{\boldsymbol{\beta}}$ and $\hat{v}_i$ in $\hat{\mathbf{v}} = (\hat{v}_1, \ldots, \hat{v}_m)^T$ satisfy

$$\sum_{i=1}^{m}\sum_{j\in s_i}\mathbf{x}_{ij}\left(y_{ij} - \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}} - \hat{v}_i\right) = 0. \tag{2.10}$$

Estimators $\hat{\boldsymbol{\beta}}$ and $\hat{v}_i$ are used to compute EBLUP predictions $\hat{y}_{ij}^{\text{EBLUP}}$ for the $N_i - n_i$ unobserved units in small area i: $\hat{y}_{ij}^{\text{EBLUP}} = \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}} + \hat{v}_i$ for $j \in r_i$. An EBLUP estimator of $\bar{Y}_i$, denoted as $\hat{\bar{Y}}_i^{\text{EBLUP}}$, is obtained by replacing in (2.4) $\hat{y}_{ij}$ by $\hat{y}_{ij}^{\text{EBLUP}}$. It follows that $\hat{\bar{Y}}_i^{\text{EBLUP}}$ is

$$\hat{\bar{Y}}_i^{\text{EBLUP}} = \frac{1}{N_i}\left[\sum_{j\in s_i} y_{ij} + \mathbf{x}_{ir}^T\hat{\boldsymbol{\beta}} + (N_i - n_i)\,\hat{v}_i\right], \tag{2.11}$$

where $\mathbf{x}_{ir} = \sum_{j \in r_i} \mathbf{x}_{ij}$ represents the sum of non sampled values $\mathbf{x}_{ij}$.

2.2 You-Rao estimation

You and Rao (2002) proposed a pseudo-EBLUP small area mean estimator (YR estimator) that incorporates the design weights d_{ij} into the formula of the EBLUP estimator. A property of the pseudo-EBLUP estimator is that the design consistency is preserved as the area sample size increases. Furthermore, the YR predictor offers protection against model failure or an informative sampling design (see among others Hidiroglou and Estevao, 2016 and Verret, Rao and Hidiroglou, 2015 for details). Pseudo EBLUP estimators can be constructed using the procedure in You and Rao (2002) with survey weights w_{ij} that may be calibrated on some vector of auxiliary variables. Let $\hat{\boldsymbol{\beta}}^{\mathrm{YR}}$ and $\hat{\mathbf{v}}^{\mathrm{YR}} = \left(\hat{v}_1^{\mathrm{YR}}, \ldots, \hat{v}_m^{\mathrm{YR}}\right)^T$ be the YR estimators of $\boldsymbol{\beta}$ and $\mathbf{v}$ respectively based on weights w_{ij} (see You and Rao, 2002 for details). The estimators $\hat{\boldsymbol{\beta}}^{\mathrm{YR}}$ and $\hat{v}_i^{\mathrm{YR}}$ satisfy the estimating unit-level based equations

$$\sum_{i=1}^{m} \sum_{j \in s_i} w_{ij} \mathbf{x}_{ij} \left(y_{ij} - \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}^{\mathrm{YR}} - \hat{v}_i^{\mathrm{YR}}\right) = 0. \tag{2.12}$$

Equations (2.12) represent the survey-weighted version of equations (2.10). You-Rao predictions $\hat{y}_{ij}^{\mathrm{YR}}$ of y_{ij} are computed as $\hat{y}_{ij}^{\mathrm{YR}} = \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}^{\mathrm{YR}} + \hat{v}_i^{\mathrm{YR}}$ for $j \in r_i$. Replacing $\hat{y}_{ij}$ by $\hat{y}_{ij}^{\mathrm{YR}}$ in (2.4) leads to the YR estimator of $\bar{Y}_i$ in the case of non negligible sampling rates:

$$\hat{\bar{Y}}_i^{\mathrm{YR}} = \frac{1}{N_i}\left[\sum_{j \in s_i} y_{ij} + \mathbf{x}_{ir}^T \hat{\boldsymbol{\beta}}^{\mathrm{YR}} + \left(N_i - n_i\right) \hat{v}_i^{\mathrm{YR}}\right]. \tag{2.13}$$

Estimators $\hat{\boldsymbol{\beta}}^{\mathrm{YR}}$ and $\hat{\mathbf{v}}^{\mathrm{YR}}$ can alternatively be obtained as solutions to weighted mixed model equations similar to (2.6) (see Huang and Hidiroglou, 2003 for details). To this end, we define matrices $\mathbf{W}_i = \mathrm{diag}_{1 \le j \le n_i}\{w_{ij}\}$, $\mathbf{W} = \mathrm{diag}_{1 \le i \le m}\{\mathbf{W}_i\}$ and $\boldsymbol{\Omega} = \mathrm{diag}_{1 \le i \le m}\{\omega_i\}$, where $\omega_i = \sum_{j \in s_i} w_{ij}^2 \Big/ \sum_{j \in s_i} w_{ij}$ for $i = 1, \ldots, m$. Let ϕ_w be the sample weighted version of ϕ, where

$$\phi_w = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{v})^T \mathbf{W}^{1/2}\mathbf{R}^{-1}\mathbf{W}^{1/2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{v}) + \mathbf{v}^T \boldsymbol{\Omega}^{1/2}\mathbf{G}^{-1}\boldsymbol{\Omega}^{1/2}\mathbf{v}, \tag{2.14}$$

with $\mathbf{W}^{1/2}$ and $\boldsymbol{\Omega}^{1/2}$ representing the square root of matrices $\mathbf{W}$ and $\boldsymbol{\Omega}$ respectively. In the first term of ϕ_w, the model error associated with the observation y_{ij} is weighted by the corresponding survey weight w_{ij}, whereas in the second term of ϕ_w, the factor ω_i in the diagonal matrix $\boldsymbol{\Omega}$ represents the weight attached to the small area effect v_i. It can be shown that the minimization of ϕ_w with respect to $\boldsymbol{\beta}$ and $\mathbf{v}$ leads to $\left(\hat{\boldsymbol{\beta}}^{\mathrm{YR}}, \hat{\mathbf{v}}^{\mathrm{YR}}\right)$. It follows that $\left(\hat{\boldsymbol{\beta}}^{\mathrm{YR}}, \hat{\mathbf{v}}^{\mathrm{YR}}\right)$ are given by

$$\begin{pmatrix} \hat{\boldsymbol{\beta}}^{\mathrm{YR}} \\ \hat{\mathbf{v}}^{\mathrm{YR}} \end{pmatrix} = \hat{\mathbf{A}}_w^{-1} \hat{\mathbf{b}}_w, \tag{2.15}$$

where the known values of $\mathbf{A}_w$ and $\mathbf{b}_w$ are given by

$$\mathbf{A}_w = \begin{pmatrix} \mathbf{X}^T\mathbf{W}^{1/2}\mathbf{R}^{-1}\mathbf{W}^{1/2}\mathbf{X} & \mathbf{X}^T\mathbf{W}^{1/2}\mathbf{R}^{-1}\mathbf{W}^{1/2}\mathbf{Z} \\ \mathbf{Z}^T\mathbf{W}^{1/2}\mathbf{R}^{-1}\mathbf{W}^{1/2}\mathbf{X} & \mathbf{Z}^T\mathbf{W}^{1/2}\mathbf{R}^{-1}\mathbf{W}^{1/2}\mathbf{Z} + \boldsymbol{\Omega}^{1/2}\mathbf{G}^{-1}\boldsymbol{\Omega}^{1/2} \end{pmatrix} \text{ and } \mathbf{b}_w = \begin{pmatrix} \mathbf{X}^T\mathbf{W}^{1/2}\mathbf{R}^{-1}\mathbf{W}^{1/2}\mathbf{y} \\ \mathbf{Z}^T\mathbf{W}^{1/2}\mathbf{R}^{-1}\mathbf{W}^{1/2}\mathbf{y} \end{pmatrix}, \tag{2.16}$$

and $\hat{\mathbf{A}}_w$ and $\hat{\mathbf{b}}_w$ are empirical versions of $\mathbf{A}_w$ and $\mathbf{b}_w$ obtained by estimating $\mathbf{G}$ and $\mathbf{R}$ by $\hat{\mathbf{G}} = \hat{\sigma}_v^2 \mathbf{I}_m$ and $\hat{\mathbf{R}} = \hat{\sigma}_e^2 \mathbf{I}_n$ respectively. Equation (2.15) can alternatively be written as

$$\begin{pmatrix} \hat{\boldsymbol{\beta}}^{\mathrm{YR}} \\ \hat{\mathbf{v}}^{\mathrm{YR}} \end{pmatrix} = \begin{pmatrix} \left(\mathbf{X}^T \hat{\mathbf{V}}_w^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^T \hat{\mathbf{V}}_w^{-1} \mathbf{y} \\ \hat{\mathbf{G}}_\omega \mathbf{Z}^T \hat{\mathbf{V}}_w^{-1} \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\mathrm{YR}}\right) \end{pmatrix}, \tag{2.17}$$

where $\hat{\mathbf{G}}_\omega = \boldsymbol{\Omega}^{-1/2} \hat{\mathbf{G}} \boldsymbol{\Omega}^{-1/2}$ and $\hat{\mathbf{V}}_w = \mathbf{W}^{-1/2} \hat{\mathbf{R}} \mathbf{W}^{-1/2} + \mathbf{Z}\hat{\mathbf{G}}_\omega \mathbf{Z}^T$.

3. Benchmarked estimators

We now proceed to develop benchmarked estimators of the small area means $\bar{Y}_i$ using unit level model (2.2) or augmented versions of it. We assume that a reliable direct estimator $\hat{Y}_w = \sum_{i=1}^m \sum_{j \in s_i} w_{ij} y_{ij}$ of the population total Y is available, where $Y = \sum_{i=1}^m Y_i$, and $Y_i = N_i \bar{Y}_i$ is the total of small area i. Let $\hat{\bar{Y}}_i$ be the model-based small area estimator of $\bar{Y}_i$. It is desirable to ensure that the aggregated values of $\hat{\bar{Y}}_i$, agree with the reliable estimator $\hat{Y}_w$. The small area means estimators $\hat{\bar{Y}}_i$, $i = 1, \ldots, m$, are said to be benchmarked to $\hat{Y}_w$ if

$$\sum_{i=1}^m N_i \hat{\bar{Y}}_i = \hat{Y}_w. \tag{3.1}$$

Let $\hat{Y}_w$ be a GREG estimator with weights calibrated at the population level on a vector of auxiliary variables $\mathbf{x}_{ij}^*$. This estimator is analogous to the combined regression estimator if one views the small areas as strata. The vector of auxiliary variables $\mathbf{x}_{ij}^*$ may or may not be the same as $\mathbf{x}_{ij}$. We distinguish two cases in this context: $\mathbf{x}_{ij} \subseteq \mathbf{x}_{ij}^*$ and $\mathbf{x}_{ij} \not\subset \mathbf{x}_{ij}^*$. The first case, $\mathbf{x}_{ij} \subseteq \mathbf{x}_{ij}^*$, implies that all the components of $\mathbf{x}_{ij}$ also belong to $\mathbf{x}_{ij}^*$, and that $\mathbf{x}_{ij}^*$ may or may not have additional components that are different from those contained in $\mathbf{x}_{ij}$. The second case, $\mathbf{x}_{ij} \not\subset \mathbf{x}_{ij}^*$, implies that some of the components of $\mathbf{x}_{ij}$ do not appear in $\mathbf{x}_{ij}^*$. We assume that the first component of both vectors $\mathbf{x}_{ij}$ and $\mathbf{x}_{ij}^*$ are equal to one, as they represent an intercept term.

For a given sample s, auxiliary data $\mathbf{x}_{ij}^*$ and basic design weights $d_{ij} = 1/\pi_{ij}$, the GREG estimator of the population total Y is given by

$$\hat{Y}^{\mathrm{GREG}} = \sum_{i=1}^m \sum_{j \in s_i} w_{ij}^{\mathrm{GREG}} y_{ij},$$

where the GREG weights w_{ij}^{GREG} are given by

$$w_{ij}^{\mathrm{GREG}} = d_{ij} \left(1 + \left(\mathbf{X}^* - \hat{\mathbf{X}}^{*\mathrm{HT}}\right)^T \left(\sum_{i=1}^m \sum_{j \in s_i} d_{ij} \mathbf{x}_{ij}^* \mathbf{x}_{ij}^{*T}\right)^{-1} \mathbf{x}_{ij}^*\right). \tag{3.2}$$

In equation (3.2), $\mathbf{X}^* = \sum_{i=1}^m \mathbf{X}_i^*$, where $\mathbf{X}_i^* = \sum_{j=1}^{N_i} \mathbf{x}_{ij}^*$ represents the known small area total, whereas $\hat{\mathbf{X}}^{*\mathrm{HT}} = \sum_{i=1}^m \hat{\mathbf{X}}_i^{*\mathrm{HT}}$ and $\hat{\mathbf{X}}_i^{*\mathrm{HT}} = \sum_{j \in s_i} d_{ij} \mathbf{x}_{ij}^*$ represent respectively the direct design-based Horvitz-Thompson estimators of $\mathbf{X}^*$ and $\mathbf{X}_i^*$. Note that

$$\sum_{i=1}^{m}\sum_{j\in s_i} w_{ij}^{\mathrm{GREG}}\mathbf{x}_{ij}^{*} = \mathbf{X}^{*}. \tag{3.3}$$

Using the GREG weights w_{ij}^{GREG}, estimators of N_i and $\mathbf{X}_i$ are given by

$$\hat{N}_i^{\mathrm{GREG}} = \sum_{j\in s_i} w_{ij}^{\mathrm{GREG}} \quad \text{and} \quad \hat{\mathbf{X}}_i^{\mathrm{GREG}} = \sum_{j\in s_i} w_{ij}^{\mathrm{GREG}}\mathbf{x}_{ij}. \tag{3.4}$$

The small area estimates $\hat{\bar{Y}}_i^{\mathrm{EBLUP}}$ and $\hat{\bar{Y}}_i^{\mathrm{YR}}$ given respectively by (2.11) and (2.13), do not satisfy the benchmarking equation (3.1) for $\hat{Y}_w = \hat{Y}^{\mathrm{GREG}}$: that is the total estimates $\hat{Y}^{\mathrm{EBLUP}} = \sum_{i=1}^{m} N_i\hat{\bar{Y}}_i^{\mathrm{EBLUP}}$, and $\hat{Y}^{\mathrm{YR}} = \sum_{i=1}^{m} N_i\hat{\bar{Y}}_i^{\mathrm{YR}}$ do not match the GREG estimator $\hat{Y}^{\mathrm{GREG}}$. We need to adjust $\hat{\bar{Y}}_i^{\mathrm{EBLUP}}$ and $\hat{\bar{Y}}_i^{\mathrm{YR}}$ so that the sum of these modified small area estimators add up to $\hat{Y}^{\mathrm{GREG}}$ when they are summed over all the m small areas.

A very simple modification to the $\hat{\bar{Y}}_i^{\mathrm{EBLUP}}$'s and $\hat{\bar{Y}}_i^{\mathrm{YR}}$'s is called ratio benchmarking. It consists of multiplying each $\hat{\bar{Y}}_i^{\mathrm{EBLUP}}$ and $\hat{\bar{Y}}_i^{\mathrm{YR}}$ by the common adjustment factors $\hat{Y}^{\mathrm{GREG}}\Big/\sum_{i=1}^{m} N_i\hat{\bar{Y}}_i^{\mathrm{EBLUP}}$ and $\hat{Y}^{\mathrm{GREG}}\Big/\sum_{i=1}^{m} N_i\hat{\bar{Y}}_i^{\mathrm{YR}}$ respectively, leading to the ratio benchmarked estimators

$$\hat{\bar{Y}}_{ib}^{\mathrm{EBRat}} = \hat{\bar{Y}}_i^{\mathrm{EBLUP}}\frac{\hat{Y}^{\mathrm{GREG}}}{\sum_{i=1}^{m} N_i\hat{\bar{Y}}_i^{\mathrm{EBLUP}}} \quad \text{and} \quad \hat{\bar{Y}}_{ib}^{\mathrm{YRat}} = \hat{\bar{Y}}_i^{\mathrm{YR}}\frac{\hat{Y}^{\mathrm{GREG}}}{\sum_{i=1}^{m} N_i\hat{\bar{Y}}_i^{\mathrm{YR}}}. \tag{3.5}$$

It readily follows that both $\hat{\bar{Y}}_{ib}^{\mathrm{EBRat}}$ and $\hat{\bar{Y}}_{ib}^{\mathrm{YRat}}$ satisfy equation (3.1) with $\hat{Y}_w = \hat{Y}^{\mathrm{GREG}}$. In equation (3.5) and hereafter the subscript b denotes that the estimators are benchmarked to $\hat{Y}^{\mathrm{GREG}}$.

Note that the $\hat{\bar{Y}}_i^{\mathrm{EBLUP}}$'s and $\hat{\bar{Y}}_i^{\mathrm{YR}}$'s in equation (3.5) are multiplied by the same factor regardless of their precision and ignoring the particular small area characteristics, such as the variability of the units within a small area, or the small area sample size. Consequently, the resulting benchmarked estimators, $\hat{\bar{Y}}_{ib}^{\mathrm{EBRat}}$ and $\hat{\bar{Y}}_{ib}^{\mathrm{YRat}}$, based on this simple procedure, are just proportional modifications of estimators $\hat{\bar{Y}}_i^{\mathrm{EBLUP}}$ and $\hat{\bar{Y}}_i^{\mathrm{YR}}$ respectively, to obtain the desired concordance. This limitation can be avoided by using the small area model (2.2) to construct the benchmarked estimators.

We now proceed to show how model (2.2) can be used to obtain estimators benchmarked to $\hat{Y}^{\mathrm{GREG}}$. In Sections 3.1 and 3.2 we adapt the procedures in Stefan and Hidiroglou (2020) for obtaining benchmarked estimators to the case of non-negligible sampling rates. In Sections 3.3 and 3.4 we introduce two restricted benchmarked estimators based on the procedure proposed by Ugarte et al. (2009). The benchmarked estimators of Sections 3.1 and 3.2 rely on the assumption that $\mathbf{x}_{ij} \subseteq \mathbf{x}_{ij}^{*}$, whereas the estimators of Sections 3.3 and 3.4 can be computed for any vector $\mathbf{x}_{ij}$ or $\mathbf{x}_{ij}^{*}$.

3.1 Augmented EBLUP benchmarked estimators

The GREG weights w_{ij}^{GREG} should be used in the estimation to achieve benchmarking to $\hat{Y}^{\mathrm{GREG}}$. A possible way that w_{ij}^{GREG} can be incorporated in the estimation is by augmenting the small area model (2.2) with a suitable auxiliary variable that is a function of w_{ij}^{GREG}. This procedure is based on the augmented model approach used by Wang et al. (2008), whereby estimates obtained using the FH

area-level model could be forced to add up to specified totals. Stefan and Hidiroglou (2020) adapted the Wang et al. (2008) approach under the basic unit-level model and for negligible sampling rates. They showed that benchmarking to $\hat{Y}^{\text{GREG}}$ could be obtained by augmenting model (2.2) with the GREG weights w_{ij}^{GREG}. We extend Stefan and Hidiroglou (2020) to the case when the sampling rates are non-negligible. For this case, benchmarking to $\hat{Y}^{\text{GREG}}$ is achieved by augmenting model (2.2) with $q_{ij} = w_{ij}^{\text{GREG}} - 1$. This leads to the augmented model given by

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta}_{1a} + q_{ij}\beta_{2a} + v_{ia} + e_{ija},\ i = 1, \ldots, m;\ j \in s_i. \tag{3.6}$$

The random effects v_{ia} are assumed to be i.i.d. $N\left(0, \sigma_{va}^2\right)$ and independent of the unit errors e_{ija}, and the e_{ija}'s are assumed to be i.i.d. $N\left(0, \sigma_{ea}^2\right)$. The EBLUP estimators of $\boldsymbol{\beta}_a = \left(\boldsymbol{\beta}_{1a}^T, \beta_{2a}\right)^T$ and v_{ia} in (3.6) are respectively denoted by $\hat{\boldsymbol{\beta}}_a = \left(\hat{\boldsymbol{\beta}}_{1a}^T, \hat{\beta}_{2a}\right)^T$ and $\hat{v}_{ia}$. We can now spell Result 1 for $\hat{\boldsymbol{\beta}}_a$ and $\hat{v}_{ia}$.

Result 1. The EBLUP estimators $\hat{\boldsymbol{\beta}}_a$ and $\hat{v}_{ia}$ based on model (3.6) obey the following equation

$$\sum_{i=1}^{m}\sum_{j\in s_i} y_{ij} + \left(\sum_{i=1}^{m} \mathbf{x}_{ir}\right)^T \hat{\boldsymbol{\beta}}_{1a} + \sum_{i=1}^{m} q_{iw}\hat{\beta}_{2a} + \sum_{i=1}^{m}\left(\hat{N}_i^{\text{GREG}} - n_i\right)\hat{v}_{ia} = \hat{Y}^{\text{GREG}}, \tag{3.7}$$

where $q_{iw} = \sum_{j\in s_i} q_{ij}^2 = \sum_{j\in s_i} (w_{ij}^{\text{GREG}} - 1)^2$.

Proof: See Appendix A.

It follows from equation (3.7) that small area estimators benchmarked to $\hat{Y}^{\text{GREG}}$ are given by

$$\hat{\bar{Y}}_{iab}^{\text{EBLUP}} = \frac{1}{N_i}\left[\sum_{j\in s_i} y_{ij} + \mathbf{x}_{ir}^T\hat{\boldsymbol{\beta}}_{1a} + q_{iw}\hat{\beta}_{2a} + \left(\hat{N}_i^{\text{GREG}} - n_i\right)\hat{v}_{ia}\right]. \tag{3.8}$$

The subscript a indicates that $\hat{\bar{Y}}_{iab}^{\text{EBLUP}}$ is based on an augmented small area model.

3.2 You-Rao benchmarked estimators

The procedure proposed by You and Rao (2002) can be used with any survey weights w_{ij}. However, there is no guarantee that the resulting YR estimator will be benchmarked to $\hat{Y}^{\text{GREG}}$. When the sampling rates are negligible, Stefan and Hidiroglou (2020) obtained benchmarked estimators with the You and Rao's (2002) procedure based on the weights $w_{ij} = w_{ij}^{\text{GREG}}$ of the GREG estimator. When the sampling rates are non-negligible, we now show that the weights $w_{ij} = w_{ij}^{\text{GREG}} - 1$ lead to YR benchmarked estimators.

Let $\hat{\boldsymbol{\beta}}^{\text{YR}}$ and $\hat{\mathbf{v}}^{\text{YR}} = \left(\hat{v}_1^{\text{YR}}, \ldots, \hat{v}_m^{\text{YR}}\right)^T$ be YR estimators of $\boldsymbol{\beta}$ and $\mathbf{v}$ respectively with w_{ij} replaced by $w_{ij}^{\text{GREG}} - 1$. Using $\hat{\boldsymbol{\beta}}^{\text{YR}}$, $\hat{v}_i^{\text{YR}}$ and the $N_i - n_i$ estimates $\hat{y}_{ij}^{\text{YR}} = \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}^{\text{YR}} + \hat{v}_i^{\text{YR}}$ for $j \in r_i$, a YR estimator, denoted as $\hat{\bar{Y}}_i^{\text{YR}}$, can be computed with equation (2.13). However, $\hat{\bar{Y}}_i^{\text{YR}}$ is not benchmarked to $\hat{Y}^{\text{GREG}}$ even if it uses the weights $w_{ij}^{\text{GREG}} - 1$. The original YR procedure leads to a self-benchmarked estimator in a limited number of cases.

To achieve the benchmark to $\hat{Y}^{\text{GREG}}$, a YR modified estimator, denoted as $\hat{\bar{Y}}_{ib}^{\text{YR}}$, is defined as follows:

$$\hat{\bar{Y}}_{ib}^{\mathrm{YR}} = \frac{1}{N_i}\left[\sum_{j\in s_i} y_{ij} + \mathbf{x}_{ir}^T\hat{\boldsymbol{\beta}}^{\mathrm{YR}} + \left(\hat{N}_i^{\mathrm{GREG}} - n_i\right)\hat{v}_i^{\mathrm{YR}}\right]. \tag{3.9}$$

The following proves that $\hat{\bar{Y}}_{ib}^{\mathrm{YR}}$ defined by (3.9) benchmarks to $\hat{Y}^{\mathrm{GREG}}$.

Result 2. Let $\hat{\boldsymbol{\beta}}^{\mathrm{YR}}$ and $\hat{\mathbf{v}}^{\mathrm{YR}} = \left(\hat{v}_1^{\mathrm{YR}}, \ldots, \hat{v}_m^{\mathrm{YR}}\right)^T$ be respectively the YR estimators of $\boldsymbol{\beta}$ and $\mathbf{v}$, constructed with weights $w_{ij}^{\mathrm{GREG}} - 1$. Then, $\left(\hat{\boldsymbol{\beta}}^{\mathrm{YR}}, \hat{\mathbf{v}}^{\mathrm{YR}}\right)$ satisfy the following equation:

$$\sum_{i=1}^{m}\sum_{j\in s_i} y_{ij} + \sum_{i=1}^{m}\mathbf{x}_{ir}^T\hat{\boldsymbol{\beta}}^{\mathrm{YR}} + \sum_{i=1}^{m}\left(\hat{N}_i^{\mathrm{GREG}} - n_i\right)\hat{v}_i^{\mathrm{YR}} = \hat{Y}^{\mathrm{GREG}}.$$

Proof: See Appendix A.

Given $\mathbf{x}_{ij}^*$, the weights w_{ij}^{GREG} are calibrated on $\mathbf{x}_{ij}^*$ at the small area level if they satisfy the following equations

$$\sum_{j\in s_i} w_{ij}^{\mathrm{GREG}}\mathbf{x}_{ij}^* = \mathbf{X}_i^*, \quad \text{for} \quad i = 1, \ldots, m. \tag{3.10}$$

Equations (3.10) implies equation (3.3), however, the reverse is not true. If the weights w_{ij}^{GREG} satisfy (3.10), and since $\mathbf{x}_{ij} \subseteq \mathbf{x}_{ij}^*$, it follows that the weights w_{ij}^{GREG} are also calibrated on $\mathbf{x}_{ij}$ at the small area level. In turn, this implies that $\hat{N}_i^{\mathrm{GREG}} = N_i$, as we assume that vector $\mathbf{x}_{ij}$ contains the constant regressor equal to 1. It follows that $\hat{\bar{Y}}_i^{\mathrm{YR}} = \hat{\bar{Y}}_{ib}^{\mathrm{YR}}$. Thus, the YR estimator $\hat{\bar{Y}}_i^{\mathrm{YR}}$ constructed with $w_{ij}^{\mathrm{GREG}} - 1$ is self-benchmarked to $\hat{Y}^{\mathrm{GREG}}$ in the special case when the GREG weights are calibrated at the small area level (see You and Rao, 2002).

3.3 Restricted EBLUP benchmarked estimator

In Section 2 we showed that the EBLUP estimators of $(\boldsymbol{\beta}, \mathbf{v})$ can be obtained if the function ϕ defined in (2.5) is minimized with respect to $(\boldsymbol{\beta}, \mathbf{v})$. It therefore follows that an EBLUP estimator can be viewed as the solution to an unrestricted minimization problem. The idea of restricted EBLUP estimators is to obtain new estimators of $(\boldsymbol{\beta}, \mathbf{v})$ by minimizing ϕ subject to the restriction given by the benchmark condition. The procedure was used by Pfeffermann and Barnard (1991) under the FH area-level model. More recently, Ugarte et al. (2009) applied the procedure under the BHF unit-level model to obtain benchmarking to a synthetic estimator. Ugarte et al. (2009) described the restricted estimator as a generalized least squares estimator subject to a restriction by noticing that the minimization can be conducted as in the econometrics theory of regression estimation under linear constraints. We now describe the procedure in Ugarte et al. (2009).

We denote by $\hat{\boldsymbol{\beta}}^R$ and $\hat{\mathbf{v}}^R = \left(\hat{v}_1^R, \ldots, \hat{v}_m^R\right)^T$ the new restricted EBLUP estimators of $(\boldsymbol{\beta}, \mathbf{v})$. Then, the restricted EBLUP estimator of $\bar{Y}_i$, denoted as $\hat{\bar{Y}}_{ib}^{\mathrm{REBLUP}}$, is given by equation (2.4), where $\hat{y}_{ij}$ are replaced by $\hat{y}_{ij}^R = \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}^R + \hat{v}_i^R$, for $j \in r_i$. We impose that the estimators $\hat{\bar{Y}}_{ib}^{\mathrm{REBLUP}}$, $i = 1, \ldots, m$ be benchmarked to $\hat{Y}^{\mathrm{GREG}}$, that is they satisfy equation (3.1) with $\hat{Y}_w = \hat{Y}^{\mathrm{GREG}}$. After carrying out some algebra, it can be shown that the benchmark to $\hat{Y}^{\mathrm{GREG}}$ of estimators $\hat{\bar{Y}}_{ib}^{\mathrm{REBLUP}}$, $i = 1, \ldots, m$ is equivalent to the following linear constraint equation

$$\mathbf{a}_1^T \hat{\boldsymbol{\beta}}^R + \mathbf{a}_2^T \hat{\mathbf{v}}^R = \hat{Y}_r^{\text{GREG}}, \tag{3.11}$$

where $\mathbf{a}_1 = \sum_{i=1}^m \mathbf{x}_{ir}$, $\mathbf{a}_2 = (N_1 - n_1, \ldots, N_m - n_m)^T$, $Y_r = Y - \sum_{i=1}^m \sum_{j \in s_i} y_{ij}$ is the total of non-observed y_{ij} values with $i = 1, \ldots, m$; $j \in r_i$, and $\hat{Y}_r^{\text{GREG}} = \hat{Y}^{\text{GREG}} - \sum_{i=1}^m \sum_{j \in s_i} y_{ij}$ is an estimator of Y_r based on $\hat{Y}^{\text{GREG}}$. The restricted EBLUP estimators $(\hat{\boldsymbol{\beta}}^R, \hat{\mathbf{v}}^R)$ are therefore obtained as the solution to the minimization of function ϕ given by (2.5) subject to the linear constraint (3.11).

The Lagrange multiplier method can be used to solve the constrained minimization of ϕ. After straightforward algebra, it can be shown that estimators $(\hat{\boldsymbol{\beta}}^R, \hat{\mathbf{v}}^R)$ are given by

$$\begin{pmatrix} \hat{\boldsymbol{\beta}}^R \\ \hat{\mathbf{v}}^R \end{pmatrix} = \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{v}} \end{pmatrix} + \frac{1}{\mathbf{a}^T \hat{\mathbf{A}} \mathbf{a}} \hat{\mathbf{A}}^{-1} \mathbf{a} \left[\hat{Y}_r^{\text{GREG}} - \mathbf{a}^T \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{v}} \end{pmatrix} \right], \tag{3.12}$$

where $(\hat{\boldsymbol{\beta}}, \hat{\mathbf{v}})$ are the (unconstrained) EBLUP estimators of $(\boldsymbol{\beta}, \mathbf{v})$, $\hat{\mathbf{A}}$ is the empirical version of matrix $\mathbf{A}$ defined in (2.7), and $\mathbf{a} = (\mathbf{a}_1^T \ \mathbf{a}_2^T)^T$. Then, using $\hat{y}_{ij}^R$ in (2.4), the estimator $\hat{\bar{Y}}_{ib}^{\text{REBLUP}}$ can be rewritten as

$$\hat{\bar{Y}}_{ib}^{\text{REBLUP}} = \frac{1}{N_i} \left[\sum_{j \in s_i} y_{ij} + \mathbf{x}_{ir}^T \hat{\boldsymbol{\beta}}^R + (N_i - n_i) \hat{v}_i^R \right]. \tag{3.13}$$

Remark 2. The matrix $\hat{\mathbf{A}}$ does not exist for samples when $\hat{\sigma}_v^2 = 0$. In such cases, we noticed that equation (2.8) cannot be used to compute the unconstrained estimators $(\hat{\boldsymbol{\beta}}, \hat{\mathbf{v}})$. However $(\hat{\boldsymbol{\beta}}, \hat{\mathbf{v}})$ can still be computed when $\hat{\sigma}_v^2 = 0$ because the alternative equation (2.9) can be used for $(\hat{\boldsymbol{\beta}}, \hat{\mathbf{v}})$. Equation (3.12) clearly shows that the constrained $(\hat{\boldsymbol{\beta}}^R, \hat{\mathbf{v}}^R)$ cannot be computed for samples when estimator $\hat{\sigma}_v^2$ is truncated to zero, and no alternative equation exists in these cases.

It, therefore, follows that the methods of estimation for the variance components commonly used in SAE cannot be used to compute the restricted EBLUP estimator. In Section 3.4 and Appendix B we describe an alternative method that produces a strictly positive estimation of σ_v^2 that can be applied in conjunction with $(\hat{\boldsymbol{\beta}}^R, \hat{\mathbf{v}}^R)$ such that a restricted benchmarked estimator of $\bar{Y}_i$ always exists.

3.4 Restricted You-Rao benchmarked estimator

We showed in Section 2.2 that YR estimators of $\boldsymbol{\beta}$ and $\mathbf{v}$ can be obtained as a solution to mixed model equations obtained by minimizing the sample weighted function ϕ_w given by (2.14). That is, we showed that, by defining a function ϕ_w with weights $\{w_{ij}\}$, $i = 1, \ldots, m$; $j \in s_i$ and $\{\omega_i\}$, $i = 1, \ldots, m$, and then minimizing ϕ_w, we obtain the same estimators as those given by the You and Rao's (2002) procedure. We now minimize function ϕ_w under the benchmark constraint given by (3.11). The result is a restricted YR estimator that is benchmarked to $\hat{Y}^{\text{GREG}}$.

Minimization of ϕ_w given the benchmark restriction (3.11) results in estimators of $\bar{Y}_i$, $i = 1, \ldots, m$ that are guaranteed to be benchmarked for any weights that define the function ϕ_w. Thus, one may choose any set of weights w_{ij} in ϕ_w. In a limited design-based simulation study, we compared three restricted YR estimators based on three options with respect to w_{ij}: i. $w_{ij} = w_{ij}^{\text{GREG}} - 1$; ii. $w_{ij} = w_{ij}^{\text{GREG}}$ and iii. $w_{ij} = d_{ij}$. We found no significant difference between these three estimators in terms of design mean

squared error. Given this last point and that the unrestricted benchmarked YR estimators described in Section 3.2 were based on $w_{ij} = w_{ij}^{\mathrm{GREG}} - 1$, we chose to define the restricted YR estimator based on these weights.

Let ϕ_w be defined in terms of $w_{ij} = w_{ij}^{\mathrm{GREG}} - 1$ and $\omega_i = \sum_{j \in s_i} w_{ij}^2 \Big/ \sum_{j \in s_i} w_{ij}$. Minimization of ϕ_w with respect to $(\boldsymbol{\beta}, \mathbf{v})$ subject to the benchmark constraint (3.11) results in the restricted YR estimators of $(\boldsymbol{\beta}, \mathbf{v})$, denoted as $\left(\hat{\boldsymbol{\beta}}^{\mathrm{RYR}}, \hat{\mathbf{v}}^{\mathrm{RYR}}\right)$. They are given by:

$$\begin{pmatrix} \hat{\boldsymbol{\beta}}^{\mathrm{RYR}} \\ \hat{\mathbf{v}}^{\mathrm{RYR}} \end{pmatrix} = \begin{pmatrix} \hat{\boldsymbol{\beta}}^{\mathrm{YR}} \\ \hat{\mathbf{v}}^{\mathrm{YR}} \end{pmatrix} + \frac{1}{\mathbf{a}^T \hat{\mathbf{A}}_w \mathbf{a}} \hat{\mathbf{A}}_w^{-1} \mathbf{a} \left[\hat{Y}_r^{\mathrm{GREG}} - \mathbf{a}^T \begin{pmatrix} \hat{\boldsymbol{\beta}}^{\mathrm{YR}} \\ \hat{\mathbf{v}}^{\mathrm{YR}} \end{pmatrix} \right], \tag{3.14}$$

where estimators $\left(\hat{\boldsymbol{\beta}}^{\mathrm{YR}}, \hat{\mathbf{v}}^{\mathrm{YR}}\right)$ are given by (2.15), and $\hat{\mathbf{A}}_w$ is the empirical version of $\mathbf{A}_w$ given by (2.16). Using $\hat{\boldsymbol{\beta}}^{\mathrm{RYR}}$ and $\hat{v}_i^{\mathrm{RYR}}$ of $\hat{\mathbf{v}}^{\mathrm{RYR}} = \left(\hat{v}_1^{\mathrm{RYR}}, \ldots, \hat{v}_m^{\mathrm{RYR}}\right)^T$, restricted YR estimates $\hat{y}_{ij}^{\mathrm{RYR}} = \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}^{\mathrm{RYR}} + \hat{v}_i^{\mathrm{RYR}}$ of unobserved y_{ij} for $j \in r_i$ are then used to compute a benchmarked restricted YR estimator:

$$\hat{\bar{Y}}_{ib}^{\mathrm{RYR}} = \frac{1}{N_i} \left[\sum_{j \in s_i} y_{ij} + \mathbf{x}_{ir}^T \hat{\boldsymbol{\beta}}^{\mathrm{RYR}} + (N_i - n_i)\, \hat{v}_i^{\mathrm{RYR}} \right]. \tag{3.15}$$

As in the case of the restricted EBLUP estimator, the estimators $\left(\hat{\boldsymbol{\beta}}^{\mathrm{RYR}}, \hat{\mathbf{v}}^{\mathrm{RYR}}\right)$ given by (3.14) do not exist if FC, ML or REML results in a truncated estimate for σ_v^2. Consequently, $\bar{Y}_i$ can only be estimated by $\hat{\bar{Y}}_{ib}^{\mathrm{RYR}}$ with a method of estimation for the variance components that always leads to strictly positive estimates for σ_v^2.

A null estimate of σ_v^2 poses no problem in computing EBLUP and YR estimators. However, we noticed that the restricted EBLUP and the restricted YR estimators cannot be computed if $\hat{\sigma}_v^2 = 0$. In order to get around this problem, we use a method proposed by Moghtased-Azar, Tehranchi and Amiri-Simkooei (2014) that guarantees that the estimator of σ_v^2 will be strictly positive. This method is based on the concept of a re-parameterized restricted maximum likelihood estimation (reREML). Their idea is to use functions whose range is the set of all positive real numbers, namely positive-valued functions (PVFs), for unknown variance components in the stochastic model instead of using variance components themselves. Their numerical results showed the successful estimation of non-negativity estimation of variance components (as positive values) as well as covariance components (as negative or positive values).

We used a Fisher-scoring algorithm to obtain iteratively the reREML estimates of the variance components of the basic unit-level model given by (2.2) (see Appendix B for details). We also carried out a small simulation and found out that for area sample sizes equal to or larger than 3, the Fisher-scoring algorithm converged in less than 15 iterations. When we only considered the samples that produced a null estimate $\hat{\sigma}_v^2 = 0$, we observed that the algorithm converged even faster (see Figure 4.1 in Section 4).

4. Simulation study

We report in this section the results of a design-based simulation study as it is in line with measures that are computed by the National Statistical Offices. A design-based study is one where a fixed finite

population is first generated using an assumed model, and then for each simulation run, a sample is drawn employing the fixed finite population. The aim of the simulation study is to evaluate the properties of the benchmarked estimators described in Section 3 in terms of design bias and design mean squared error. We considered two scenarios: *Scenario 1* corresponds to the case of correct modeling, whereas *Scenario 2* corresponds to the case of incorrect modelling. Model diagnostics such as those given in Rao and Molina (2015, pages 114-118), can be used to test whether the models are correct or not. Such model diagnostics include residual analysis to detect departures from the assumed model, selection of auxiliary variables for the model, and case-deletion diagnostics to detect influential observations.

4.1 Simulation set-up for generating the finite populations

For each scenario, we considered five populations. Each population had $m = 30$ small areas, with $N_i = 100$ population units within each small area. The populations corresponding to *Scenario 1* were created using the following model

$$y_{ij} = \beta_0 + x_{ij}\beta_1 + v_i + e_{ij},\ i = 1, \ldots, m;\ j = 1, \ldots, N_i, \tag{4.1}$$

where $\beta_0 = 10$ and $\beta_1 = 5$. For generating the populations in *Scenario 2*, we split the 30 small areas into three equal groups of small areas, denoted as G_ℓ, for $\ell = 1, 2, 3$. The first group G_1 contains areas $i = 1, \ldots, 10$, the second group G_2 contains areas $i = 11, \ldots, 20$, and the third group G_3 contains areas $i = 21, \ldots, 30$. The model within a given group is given by

$$y_{ij} = \beta_{0,\ell} + x_{ij}\beta_{1,\ell} + v_i + e_{ij},\ i \in G_\ell;\ j = 1, \ldots, N_i, \tag{4.2}$$

where $(\beta_{0,1} = 10,\ \beta_{1,1} = 1)$ for areas $i \in G_1$, $(\beta_{0,2} = 20,\ \beta_{1,2} = 5)$ for areas $i \in G_2$, and $(\beta_{0,3} = 30,\ \beta_{1,3} = 10)$ for areas $i \in G_3$. Both (4.1) and (4.2) use the auxiliary variable $\mathbf{x}_{ij} = (1,\ x_{ij})^T$ whose values $x_{ij},\ j = 1, \ldots, N_i$ were generated from an exponential distribution with mean equal to 5 and variance equal to 25.

The random components in (4.1) and (4.2) were generated from the normal distributions $v_i \sim N(0, \sigma_v^2)$ and $e_{ij} \sim N(0, \sigma_e^2)$. The five populations corresponding to *Scenario 1*, denoted as A1, B1, C1, D1 and E1, were generated based on (4.1) and the following variance parameters doublets: i. $(\sigma_{Av}^2 = 0.2,\ \sigma_e^2 = 20)$ for population A1; ii. $(\sigma_{Bv}^2 = 1,\ \sigma_e^2 = 20)$ for population B1; iii. $(\sigma_{Cv}^2 = 2,\ \sigma_e^2 = 20)$ for population C1; iv. $(\sigma_{Dv}^2 = 4,\ \sigma_e^2 = 20)$ for population D1; and $(\sigma_{Ev}^2 = 20,\ \sigma_e^2 = 20)$ for population E1. Note that, for populations A1 through E1, the value of σ_e^2 is kept fixed, whereas the values for σ_v^2 vary. The σ_v^2's are chosen to obtain the following variance ratios $\delta = \sigma_v^2 / \sigma_e^2$ as 0.01, 0.05, 0.1, 0.2 and 1. The five populations in *Scenario 2*, denoted as A2, B2, C2, D2 and E2, were generated based on (4.2) with the same variance parameters doublets as for *Scenario 1*.

A stratified sampling design was used by drawing independent probability proportional to size samples (pps) of size n_i within the i^{th} small area. The small area sample sizes were taken $n_i = 3$ for $i = 1, \ldots, m$. The selection probabilities were computed as $p_{ij} = b_{ij} \Big/ \sum_{j=1}^{N_i} b_{ij}$, where the size measures

are $b_{ij} = x_{ij}$. We used Conditional Poisson Sampling (CPS) to select the pps samples within each small area (see Tillé (2006), Chapter 5). The basic design weights are given by $d_{ij} = 1/(n_i p_{ij})$.

In *Scenario 1*, we fitted the nested regression model (4.1) and its augmented version to pps sampling data selected from one of the five populations generated with model (4.1). This scenario represents correct modeling as the model fitted and the model used to generate the finite population coincide. In *Scenario 2*, we fitted the nested regression model (4.1) and its augmented version to pps sampling data selected from one of the five populations generated with model (4.2). This scenario represents incorrect modeling as the model fitted and the model used to generate the finite population do not coincide.

We selected $G =$ 30,000 stratified pps samples from each of the ten finite populations: populations A1 to E1 corresponding to *Scenario 1*, and populations A2 to E2 corresponding to *Scenario 2*. For $g = 1, \ldots, G$ let $(\hat{\sigma}_v^{2\mathrm{RE}(g)}, \hat{\sigma}_e^{2\mathrm{RE}(g)})$ and $(\hat{\sigma}_v^{2\mathrm{reRE}(g)}, \hat{\sigma}_e^{2\mathrm{reRE}(g)})$ denote respectively the estimates of (σ_v^2, σ_e^2) given by the truncated REML method and its re-parameterized version, that correspond to the g^{th} sample. The starting values in equation (B.2) were $\alpha_1^{(0)} = \log(0.1 + \hat{\sigma}_v^{2\mathrm{RE}(g)})$ and $\alpha_2^{(0)} = \log(\hat{\sigma}_e^{2\mathrm{RE}(g)})$. Equation (B.2) reached convergence in less than 15 iterations for all the populations and both scenarios. Based on the G simulated samples selected in each of the five populations corresponding to *Scenario 1*, we computed the Monte Carlo value of the probability to obtain a zero truncated REML estimate for σ_v^2 as

$$P_{\mathrm{MC}}(\hat{\sigma}_v^{2\mathrm{RE}} = 0) = \frac{1}{G}\sum_{g=1}^{G} I(\hat{\sigma}_v^{2\mathrm{RE}(g)} = 0),$$

where $I(A)$ is an indicator function with value 1 if condition A holds, and 0 otherwise.

Table 4.1 displays the Monte Carlo values of the probability to get a zero estimate for $\hat{\sigma}_v^{2\mathrm{RE}}$. It can be seen that the simulated probability $P_{\mathrm{MC}}(\hat{\sigma}_v^{2\mathrm{RE}} = 0)$ can be as high as 0.47 for $\delta = 0.01$. As δ increases, this empirical probability decreases. Table 4.1 clearly shows that estimates $(\hat{\sigma}_v^{2\mathrm{RE}}, \hat{\sigma}_e^{2\mathrm{RE}})$ cannot be used to compute the restricted EBLUP and YR estimators for samples selected in populations A1, B1, C1 and D1.

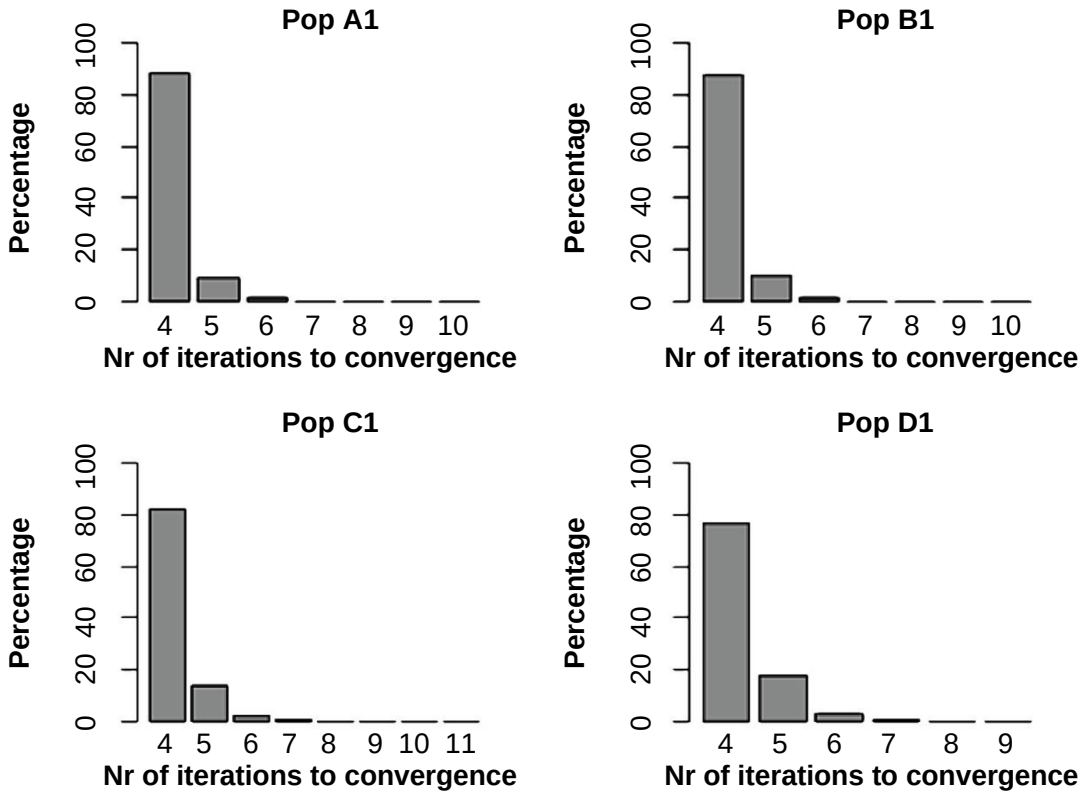
Table 4.1
Values of $P(\hat{\sigma}_v^{2\mathrm{RE}} = 0)$: *Scenario 1*

	Pop A1 $\delta = 0.01$	**Pop B1** $\delta = 0.05$	**Pop C1** $\delta = 0.1$	**Pop D1** $\delta = 0.2$	**Pop E1** $\delta = 1$
$P_{\mathrm{MC}}(\hat{\sigma}_v^{2\mathrm{RE}} = 0)$	0.47	0.40	0.21	0.06	0.00

Figure 4.1 displays the number of iterations to convergence of the Fisher-scoring algorithm for the estimate $\hat{\sigma}_v^{2\mathrm{reRE}}$ of σ_v^2. The algorithm stops when the value of $|\hat{\sigma}_v^{2\mathrm{reRE}(r+1)} - \hat{\sigma}_v^{2\mathrm{reRE}(r)}|$ is less than 10^{-5}, where $\hat{\sigma}_v^{2\mathrm{reRE}(r)}$ represents the r^{th} iteration computed with equation (B.2) in Appendix B. The percentages of Figure 4.1 are based only on samples with a truncated REML estimate of σ_v^2, that is $\hat{\sigma}_v^{2\mathrm{RE}} = 0$. We only considered populations A1, B1, C1 and D1, as these four populations have non-negligible

probabilities for $\hat{\sigma}_v^{2\text{RE}}$ to be null. Figure 4.1 clearly shows that the convergence is attained in a maximum of 11 iterations.

Figure 4.1 Percentage of iterations to convergence in samples with $\hat{\sigma}_v^{2\text{RE}} = 0$.



4.2 Comparison between the benchmarked estimators

The aim of the simulation study is to compare the benchmarked estimators described in Section 3 in terms of design bias and design mean squared error. We used both scenarios as we wanted to check how benchmarking protects against incorrect modeling. Furthermore, we considered the benchmark to two GREG estimators: $\hat{Y}_1^{\text{GREG}}$ and $\hat{Y}_2^{\text{GREG}}$. Estimator $\hat{Y}_1^{\text{GREG}}$ has weights given by (3.2) calibrated on the auxiliary vector $\mathbf{x}_{ij} = (1, x_{ij})$ associated with the small area model. It follows that estimator $\hat{Y}_1^{\text{GREG}}$ corresponds to the case $\mathbf{x}_{ij} \subseteq \mathbf{x}_{ij}^*$. The second GREG estimator $\hat{Y}_2^{\text{GREG}}$ has weights given by (3.2) based on auxiliary vector $\mathbf{x}_{ij}^* = (1, x_{ij}^*)$, where the values x_{ij}^*, $j = 1, \ldots, N_i$ were generated from an exponential distribution with mean equal to 5 and variance equal to 25, and independently of the values x_{ij}, $j = 1, \ldots, N_i$. It follows that estimator $\hat{Y}_2^{\text{GREG}}$ corresponds to the case $\mathbf{x}_{ij} \not\subset \mathbf{x}_{ij}^*$, since the auxiliary variable x_{ij} associated with the unit-level model (4.1) do not belong to vector $\mathbf{x}_{ij}^*$ used to obtain the weights associated with $\hat{Y}_2^{\text{GREG}}$.

For a fixed finite population, let $\bar{Y}_i$ be the mean of the small area i and $\hat{\bar{Y}}_i$ a generic estimator of $\bar{Y}_i$. We denote by $\hat{\bar{Y}}_i^{(g)}$ the value of $\hat{\bar{Y}}_i$ based on the g^{th} simulated sample, for $g = 1, \ldots, G$. The estimators described in Section 3 respect the benchmark property regardless of the method used to estimate the variance components. Since the restricted benchmarked estimators are based on estimates $\left(\hat{\sigma}_v^{2\text{reRE}(g)}, \hat{\sigma}_e^{2\text{reRE}(g)}\right)$, we decided to use reREML for computing $\hat{\bar{Y}}_i^{(g)}$ for each estimator $\hat{\bar{Y}}_i$ evaluated in this simulation study.

We considered the following performance measures:

Average Absolute Relative Bias

$$\overline{\text{ARB}} = \frac{1}{m}\sum_{i=1}^{m}\text{ARB}_i \quad \text{with} \quad \text{ARB}_i = \left|\frac{1}{G}\sum_{g=1}^{G}\frac{\hat{\bar{Y}}_i^{(g)}}{\bar{Y}_i} - 1\right|$$

Average Relative Root Mean Squared Error

$$\overline{\text{RRMSE}} = \frac{1}{m}\sum_{i=1}^{m}\text{RRMSE}_i \quad \text{with} \quad \text{RRMSE}_i = \sqrt{\frac{1}{G}\sum_{g=1}^{G}\left(\frac{\hat{\bar{Y}}_i^{(g)}}{\bar{Y}_i} - 1\right)^2}.$$

This portion of the simulation is summarized in four tables. We provide the results separately for *Scenarios 1* and *2*. The results for the case when the benchmarking is to $\hat{Y}_1^{\text{GREG}}$ (the case $\mathbf{x}_{ij} \subseteq \mathbf{x}_{ij}^*$) are summarized in Tables 4.2 (*Scenario 1*) and 4.3 (*Scenario 2*). Those for the case when the benchmarking is to $\hat{Y}_2^{\text{GREG}}$ (the case $\mathbf{x}_{ij} \not\subset \mathbf{x}_{ij}^*$) are summarized in Tables 4.4 (*Scenario 1*) and 4.5 (*Scenario 2*).

Benchmarking to $\hat{Y}_1^{\text{GREG}}$ *(the case* $\mathbf{x}_{ij} \subseteq \mathbf{x}_{ij}^*$*)*

We computed the $\overline{\text{ARB}}$ and $\overline{\text{RRMSE}}$ for two non benchmarked estimators, $\hat{\bar{Y}}_i^{\text{EBLUP}}$ and $\hat{\bar{Y}}_i^{\text{YR}}$, as well as their corresponding estimators benchmarked to $\hat{Y}_1^{\text{GREG}}$. For $\hat{\bar{Y}}_i^{\text{EBLUP}}$, we have three benchmarked estimators $\hat{\bar{Y}}_{ib}^{\text{EBRat}}$, $\hat{\bar{Y}}_{iab}^{\text{EBLUP}}$ and $\hat{\bar{Y}}_{ib}^{\text{REBLUP}}$, given respectively by equations (3.5), (3.8) and (3.13). For $\hat{\bar{Y}}_i^{\text{YR}}$, the corresponding benchmarked estimators are $\hat{\bar{Y}}_{ib}^{\text{YRat}}$, $\hat{\bar{Y}}_{ib}^{\text{YR}}$ and $\hat{\bar{Y}}_{ib}^{\text{RYR}}$, given respectively by equations (3.5), (3.9) and (3.15).

We first discuss their properties when the model is correct (*Scenario 1*). Comparing the $\overline{\text{ARB}}$'s across all the estimators in Table 4.2, we observe that there is not much difference between the estimators. The EBLUP estimators have somewhat smaller $\overline{\text{ARB}}$'s than the estimators based on the YR procedure. The benchmarked estimator $\hat{\bar{Y}}_{iab}^{\text{EBLUP}}$ has the smallest $\overline{\text{ARB}}$'s, whereas the $\overline{\text{ARB}}$'s of the benchmarked estimators $\hat{\bar{Y}}_{ib}^{\text{EBRat}}$ and $\hat{\bar{Y}}_{ib}^{\text{REBLUP}}$ are identical to those of $\hat{\bar{Y}}_i^{\text{EBLUP}}$. The $\overline{\text{ARB}}$ values associated with estimators $\hat{\bar{Y}}_i^{\text{YR}}$, $\hat{\bar{Y}}_{ib}^{\text{YRat}}$ and $\hat{\bar{Y}}_{ib}^{\text{RYR}}$ are close, whereas estimator $\hat{\bar{Y}}_{ib}^{\text{YR}}$ has a somewhat larger relative bias, especially for larger values of $\delta = \sigma_v^2/\sigma_e^2$. For all the estimators, the $\overline{\text{ARB}}$'s increase as δ increases: slight exceptions occur when $\delta = 1$.

Next, we report on the $\overline{\text{RRMSE}}$'s. As expected, the smallest $\overline{\text{RRMSE}}$'s are associated with $\hat{\bar{Y}}_i^{\text{EBLUP}}$, whereas estimator $\hat{\bar{Y}}_i^{\text{YR}}$ has somewhat larger $\overline{\text{RRMSE}}$ values due to the use of survey weights under

correct modeling. Benchmarking results in an increase of the $\overline{\text{RRMSE}}$. Note that the $\overline{\text{RRMSE}}$'s of the benchmarked estimators $\hat{\bar{Y}}_{iab}^{\text{EBLUP}}$ and $\hat{\bar{Y}}_{ib}^{\text{YR}}$ given in Sections 3.1 and 3.2 respectively, are higher than those associated with the restricted methods $\hat{\bar{Y}}_{ib}^{\text{REBLUP}}$ and $\hat{\bar{Y}}_{ib}^{\text{RYR}}$ given in Sections 3.3 and 3.4 respectively. The naïve ratio procedures $\hat{\bar{Y}}_{ib}^{\text{EBRat}}$ and $\hat{\bar{Y}}_{ib}^{\text{YRat}}$ have $\overline{\text{RRMSE}}$'s that are quite comparable to those of the benchmarked that use the restricted methods. The $\overline{\text{RRMSE}}$'s increase as δ increases.

We conclude the following in the case $\mathbf{x}_{ij} \subseteq \mathbf{x}_{ij}^{*}$ and when the small area model is correctly specified. The restricted benchmarked or ratio type estimators perform better than those that use an augmented model for EBLUP or a modified YR method. When the restricted or the ratio benchmarking techniques are used, the resulting estimators have bias values that are similar to those associated with their non benchmarked versions, whereas their mean squared error values are slightly larger than those of the non benchmarked versions. The small area estimators and the GREG estimator $\hat{Y}_1^{\text{GREG}}$ are based on the same auxiliary variables, whereas the model is correct. Consequently, $\hat{\bar{Y}}_i^{\text{EBLUP}}$ and $\hat{\bar{Y}}_i^{\text{YR}}$ do not have to be severely modified to achieve benchmarking to $\hat{Y}_1^{\text{GREG}}$.

Table 4.2
$\overline{\text{ARB}}$ (%) and $\overline{\text{RRMSE}}$ (%) for *Scenario 1*: the benchmark to $\hat{Y}_1^{\text{GREG}}$ $\left(\mathbf{x}_{ij} \subseteq \mathbf{x}_{ij}^{*}\right)$

Estimator	Measure	Pop A1 $\delta = 0.01$	Pop B1 $\delta = 0.05$	Pop C1 $\delta = 0.1$	Pop D1 $\delta = 0.2$	Pop E1 $\delta = 1$
$\hat{\bar{Y}}_i^{\text{EBLUP}}$	$\overline{\text{ARB}}$	1.1	1.9	2.3	2.7	2.6
	$\overline{\text{RRMSE}}$	2.7	3.4	3.9	4.9	6.5
$\hat{\bar{Y}}_i^{\text{YR}}$	$\overline{\text{ARB}}$	1.2	2.0	2.4	2.9	3.1
	$\overline{\text{RRMSE}}$	3.1	3.7	4.2	5.3	7.2
$\hat{\bar{Y}}_{ib}^{\text{EBRat}}$	$\overline{\text{ARB}}$	1.1	1.9	2.3	2.7	2.6
	$\overline{\text{RRMSE}}$	3.2	3.8	4.3	5.2	6.9
$\hat{\bar{Y}}_{ib}^{\text{YRat}}$	$\overline{\text{ARB}}$	1.2	2.0	2.4	2.9	3.1
	$\overline{\text{RRMSE}}$	3.1	3.7	4.3	5.3	7.4
$\hat{\bar{Y}}_{iab}^{\text{EBLUP}}$	$\overline{\text{ARB}}$	1.0	1.6	2.1	2.4	2.3
	$\overline{\text{RRMSE}}$	9.6	9.8	10.1	11.1	13.9
$\hat{\bar{Y}}_{ib}^{\text{YR}}$	$\overline{\text{ARB}}$	1.2	2.0	2.5	3.0	3.7
	$\overline{\text{RRMSE}}$	3.5	4.8	5.4	11.7	14.5
$\hat{\bar{Y}}_{ib}^{\text{REBLUP}}$	$\overline{\text{ARB}}$	1.1	1.9	2.3	2.7	2.6
	$\overline{\text{RRMSE}}$	3.2	3.8	4.3	5.3	7.0
$\hat{\bar{Y}}_{ib}^{\text{RYR}}$	$\overline{\text{ARB}}$	1.2	2.0	2.4	2.9	3.2
	$\overline{\text{RRMSE}}$	3.1	3.7	4.3	5.3	7.5

The results for not using the correct model are given in Table 4.3. The value of δ does not have much impact on the $\overline{\text{ARB}}$'s and $\overline{\text{RRMSE}}$'s across all estimators. The $\overline{\text{ARB}}$'s and $\overline{\text{RRMSE}}$'s of the EBLUP estimators, whether they are benchmarked or not, are higher than those associated with the YR estimators. It follows that if we have incorrect modeling, the use of the YR estimators is recommended. Since $\hat{Y}_1^{\text{GREG}}$

and the estimators based on the YR procedure use the same vector of auxiliary information, it follows that there is not much difference in terms of $\overline{\text{ARB}}$ and $\overline{\text{RRMSE}}$ between the non benchmarked estimator $\hat{\bar{Y}}_i^{\text{YR}}$ and its benchmarked versions, $\hat{\bar{Y}}_{ib}^{\text{YRat}}$, $\hat{\bar{Y}}_{ib}^{\text{YR}}$ and $\hat{\bar{Y}}_{ib}^{\text{RYR}}$. However, it can be noticed that the benchmarked estimator $\hat{\bar{Y}}_{ib}^{\text{YR}}$ has the smallest $\overline{\text{ARB}}$ values, whereas the restricted benchmarked estimator $\hat{\bar{Y}}_{ib}^{\text{RYR}}$ has the smallest $\overline{\text{RRMSE}}$'s.

Table 4.3
$\overline{\text{ARB}}$ (%) and $\overline{\text{RRMSE}}$ (%) for *Scenario 2*: the benchmark to $\hat{Y}_1^{\text{GREG}}$ $\left(\mathbf{x}_{ij} \subseteq \mathbf{x}_{ij}^*\right)$

Estimator	Measure	Pop A2 $\delta = 0.01$	Pop B2 $\delta = 0.05$	Pop C2 $\delta = 0.1$	Pop D2 $\delta = 0.2$	Pop E2 $\delta = 1$
$\hat{\bar{Y}}_i^{\text{EBLUP}}$	$\overline{\text{ARB}}$	42.3	42.7	43.2	43.0	41.5
	$\overline{\text{RRMSE}}$	59.8	60.5	61.1	60.6	59.0
$\hat{\bar{Y}}_i^{\text{YR}}$	$\overline{\text{ARB}}$	13.5	13.8	13.8	13.6	13.5
	$\overline{\text{RRMSE}}$	42.8	43.2	43.5	43.2	42.4
$\hat{\bar{Y}}_{ib}^{\text{EBRat}}$	$\overline{\text{ARB}}$	42.9	43.4	43.9	43.6	42.1
	$\overline{\text{RRMSE}}$	61.2	61.9	62.7	62.1	60.3
$\hat{\bar{Y}}_{ib}^{\text{YRat}}$	$\overline{\text{ARB}}$	13.8	14.1	14.1	13.9	13.8
	$\overline{\text{RRMSE}}$	43.9	44.4	44.7	44.4	43.5
$\hat{\bar{Y}}_{iab}^{\text{EBLUP}}$	$\overline{\text{ARB}}$	19.8	20.2	20.2	20.2	19.6
	$\overline{\text{RRMSE}}$	66.2	66.7	67.6	67.3	66.6
$\hat{\bar{Y}}_{ib}^{\text{YR}}$	$\overline{\text{ARB}}$	10.9	10.6	11.5	12.5	10.7
	$\overline{\text{RRMSE}}$	47.3	47.6	48.1	47.9	47.8
$\hat{\bar{Y}}_{ib}^{\text{REBLUP}}$	$\overline{\text{ARB}}$	41.2	41.8	41.8	41.7	40.6
	$\overline{\text{RRMSE}}$	58.2	59.0	59.1	58.9	57.4
$\hat{\bar{Y}}_{ib}^{\text{RYR}}$	$\overline{\text{ARB}}$	12.5	12.7	12.6	12.5	12.5
	$\overline{\text{RRMSE}}$	42.4	42.9	43.1	42.9	42.1

Benchmarking to $\hat{Y}_2^{\text{GREG}}$ (the case $\mathbf{x}_{ij} \not\subset \mathbf{x}_{ij}^*$)

The results of this case are given in Tables 4.4 and 4.5 for *Scenarios 1* and *2*, respectively. The weighting is with respect to w_{ij}^{GREG} given by (3.2) .We investigated the following four estimators ($\hat{\bar{Y}}_{ib}^{\text{EBRat}}$, $\hat{\bar{Y}}_{ib}^{\text{REBLUP}}$, $\hat{\bar{Y}}_{ib}^{\text{YRat}}$, and $\hat{\bar{Y}}_{ib}^{\text{RYR}}$) that are benchmarked to $\hat{Y}_2^{\text{GREG}}$. The first two estimators, $\hat{\bar{Y}}_{ib}^{\text{EBRat}}$ and $\hat{\bar{Y}}_{ib}^{\text{REBLUP}}$, are given by equations (3.5) and (3.13) respectively, while the last two, $\hat{\bar{Y}}_{ib}^{\text{YRat}}$ and $\hat{\bar{Y}}_{ib}^{\text{RYR}}$, are given by equations (3.5) and (3.15).

In Table 4.4, we summarize the average ARB and RRMSE values when the model is correct. That is, both the sample and the population data respect model (4.1). We first discuss their properties in terms of the $\overline{\text{ARB}}$'s. Comparing the $\overline{\text{ARB}}$'s across all the estimators in Table 4.4, we observe once more that, under correct modeling, the original EBLUP estimator, $\hat{\bar{Y}}_i^{\text{EBLUP}}$, has the smallest $\overline{\text{ARB}}$'s. The $\overline{\text{ARB}}$'s increase when benchmarking is required, and this is different from what we noticed from Table 4.2. There is not much difference in terms of $\overline{\text{ARB}}$ between the benchmarked estimators obtained using ratio

adjustment methods, $\hat{\bar{Y}}_{ib}^{\text{EBRat}}$ and $\hat{\bar{Y}}_{ib}^{\text{YRat}}$, and those obtained by restricted methods, $\hat{\bar{Y}}_{ib}^{\text{REBLUP}}$ and $\hat{\bar{Y}}_{ib}^{\text{RYR}}$. The $\overline{\text{ARB}}$'s increase as δ increases: slight exceptions occur when $\delta = 1$.

Next, we report on the $\overline{\text{RRMSE}}$'s. As expected, the smallest $\overline{\text{RRMSE}}$'s are associated with $\hat{\bar{Y}}_i^{\text{EBLUP}}$ which is optimal under correct modeling. Benchmarking results in an increase of $\overline{\text{RRMSE}}$. Note that the $\overline{\text{RRMSE}}$'s associated with all four benchmarking procedures in Table 4.4 are quite high compared to the $\overline{\text{RRMSE}}$'s associated with the non benchmarked estimators $\hat{\bar{Y}}_i^{\text{EBLUP}}$ and $\hat{\bar{Y}}_i^{\text{YR}}$. The estimators $\hat{\bar{Y}}_{ib}^{\text{EBRat}}$ and $\hat{\bar{Y}}_{ib}^{\text{YRat}}$ have similar efficiency, whereas $\hat{\bar{Y}}_{ib}^{\text{REBLUP}}$ and $\hat{\bar{Y}}_{ib}^{\text{RYR}}$ have $\overline{\text{RRMSE}}$ values that are somewhat larger than those of $\hat{\bar{Y}}_{ib}^{\text{EBRat}}$ and $\hat{\bar{Y}}_{ib}^{\text{YRat}}$. The $\overline{\text{RRMSE}}$'s increase as δ increases.

When $\mathbf{x}_{ij} \not\subset \mathbf{x}_{ij}^*$, there are larger differences between the small area estimators based on model (2.2) that uses the vector $\mathbf{x}_{ij}$, and the GREG estimator that uses $\mathbf{x}_{ij}^*$. Notice that we considered a somewhat extreme situation when $\mathbf{x}_{ij}$ and $\mathbf{x}_{ij}^*$ have no variable in common. It follows that the modifications needed to obtain benchmarked estimators are more accentuated in this case as compared to the case $\mathbf{x}_{ij} \subseteq \mathbf{x}_{ij}^*$. This explains why in Table 4.4 the benchmarked estimators have significantly larger $\overline{\text{ARB}}$ and $\overline{\text{RRMSE}}$ values than the estimators that are not benchmarked to $\hat{Y}_2^{\text{GREG}}$.

Table 4.4
$\overline{\text{ARB}}$ (%) and $\overline{\text{RRMSE}}$ (%) for *Scenario 1*: the benchmark to $\hat{Y}_2^{\text{GREG}}$ $\left(\mathbf{x}_{ij} \not\subset \mathbf{x}_{ij}^*\right)$

Estimator	Measure	Pop A1 $\delta = 0.01$	Pop B1 $\delta = 0.05$	Pop C1 $\delta = 0.1$	Pop D1 $\delta = 0.2$	Pop E1 $\delta = 1$
$\hat{\bar{Y}}_i^{\text{EBLUP}}$	$\overline{\text{ARB}}$	1.1	1.9	2.3	2.7	2.6
	$\overline{\text{RRMSE}}$	2.7	3.4	3.9	4.9	6.5
$\hat{\bar{Y}}_i^{\text{YR}}$	$\overline{\text{ARB}}$	1.2	2.0	2.4	2.9	3.1
	$\overline{\text{RRMSE}}$	3.1	3.7	4.2	5.3	7.2
$\hat{\bar{Y}}_{ib}^{\text{EBRat}}$	$\overline{\text{ARB}}$	4.2	4.3	4.5	4.9	4.6
	$\overline{\text{RRMSE}}$	13.0	13.2	13.5	14.0	14.6
$\hat{\bar{Y}}_{ib}^{\text{YRat}}$	$\overline{\text{ARB}}$	4.2	4.3	4.5	5.0	4.8
	$\overline{\text{RRMSE}}$	13.0	13.2	13.5	14.0	14.0
$\hat{\bar{Y}}_{ib}^{\text{REBLUP}}$	$\overline{\text{ARB}}$	4.2	4.3	4.5	5.0	4.8
	$\overline{\text{RRMSE}}$	13.1	13.3	13.5	14.1	15.0
$\hat{\bar{Y}}_{ib}^{\text{RYR}}$	$\overline{\text{ARB}}$	4.2	4.3	4.6	5.1	5.0
	$\overline{\text{RRMSE}}$	13.5	13.7	13.8	14.5	16.2

The impact of using an incorrect model is given in Table 4.5. We see that $\hat{\bar{Y}}_i^{\text{EBLUP}}$ suffers the most in terms of both $\overline{\text{ARB}}$ and $\overline{\text{RRMSE}}$ because the EBLUP procedure assumes that the model is correct. The benchmarked versions of EBLUP, $\hat{\bar{Y}}_{ib}^{\text{EBRat}}$ and $\hat{\bar{Y}}_{ib}^{\text{REBLUP}}$, also have high $\overline{\text{ARB}}$'s and $\overline{\text{RRMSE}}$'s. Although the original You and Rao (2002) estimator, $\hat{\bar{Y}}_i^{\text{YR}}$, has much smaller $\overline{\text{ARB}}$ than the EBLUP estimator, its $\overline{\text{RRMSE}}$ is fairly high. The $\overline{\text{ARB}}$ and $\overline{\text{RRMSE}}$ associated with the ratio benchmarked version of $\hat{\bar{Y}}_i^{\text{YR}}$, $\hat{\bar{Y}}_{ib}^{\text{YRat}}$, are a bit higher than those associated with $\hat{\bar{Y}}_i^{\text{YR}}$. The benchmarked YR estimator, $\hat{\bar{Y}}_{ib}^{\text{RYR}}$, which is based on the restricted procedure given in Section 3.4, has an $\overline{\text{ARB}}$ that is the smallest amongst the

estimators in Table 4.5. Due to benchmarking, its $\overline{\text{RRMSE}}$ is slightly larger than the one associated with $\hat{\bar{Y}}_i^{\text{YR}}$.

Table 4.5
$\overline{\text{ARB}}$ (%) and $\overline{\text{RRMSE}}$ (%) for *Scenario 2*: the benchmark to $\hat{Y}_2^{\text{GREG}}\left(\mathbf{x}_{ij} \not\subset \mathbf{x}_{ij}^*\right)$

Estimator	Measure	Pop A2 $\delta = 0.01$	Pop B2 $\delta = 0.05$	Pop C2 $\delta = 0.1$	Pop D2 $\delta = 0.2$	Pop E2 $\delta = 1$
$\hat{\bar{Y}}_i^{\text{EBLUP}}$	$\overline{\text{ARB}}$	42.3	42.6	43.2	43.0	41.6
	$\overline{\text{RRMSE}}$	59.8	60.4	61.1	60.7	59.1
$\hat{\bar{Y}}_i^{\text{YR}}$	$\overline{\text{ARB}}$	13.6	13.6	13.9	13.7	13.5
	$\overline{\text{RRMSE}}$	42.8	43.1	43.5	43.3	42.4
$\hat{\bar{Y}}_{ib}^{\text{EBRat}}$	$\overline{\text{ARB}}$	43.8	44.4	44.9	44.6	43.3
	$\overline{\text{RRMSE}}$	65.4	66.1	67.0	66.4	64.5
$\hat{\bar{Y}}_{ib}^{\text{YRat}}$	$\overline{\text{ARB}}$	15.0	15.2	15.6	15.2	14.9
	$\overline{\text{RRMSE}}$	47.9	48.2	48.7	48.3	47.3
$\hat{\bar{Y}}_{ib}^{\text{REBLUP}}$	$\overline{\text{ARB}}$	37.3	38.0	38.1	37.8	37.1
	$\overline{\text{RRMSE}}$	57.4	58.2	58.5	58.2	56.7
$\hat{\bar{Y}}_{ib}^{\text{RYR}}$	$\overline{\text{ARB}}$	9.9	10.1	10.4	10.0	10.1
	$\overline{\text{RRMSE}}$	43.4	43.8	44.2	43.9	43.1

5. Real data example

In this section, we compare the benchmarked estimators through a real data analysis. The data set we studied is the corn and soybean data provided by Battese et al. (1988). They considered the estimation of mean hectares of corn and soybeans per segment for twelve counties in north-central Iowa. The response variable y_{ij} is the number of hectares of corn in the j^{th} segment of the i^{th} county. The auxiliary variables, x_{1ij} and x_{2ij}, are the number of pixels classified as corn and soybeans respectively, in the j^{th} segment of the i^{th} county. We report only results for $\bar{Y}_i$, the mean number of hectares of corn per segment for county i.

Following Battese et al. (1988), we deleted the sample data from the second sample segment in Hardin county because the corn area for that segment looked erroneous. Among the twelve counties, there were three counties with a single sample segment. Following Prasad and Rao (1990), we combined these three counties into a single one, resulting in 10 counties in our data set with sample size n_i ranging from 2 to 5 in each county. The total number of segments N_i (population size) within each county ranged from 402 to 1,505. Following You and Rao (2002), we assumed simple random sampling within each county, and the basic design weight was computed as $d_{ij} = N_i / n_i$ for unit j in the i^{th} county.

We base our calculations on the unit level sampling model given by

$$y_{ij} = \beta_0 + x_{1ij}\beta_1 + x_{2ij}\beta_2 + v_i + e_{ij},\ j = 1, \ldots, n_i;\ i = 1, \ldots, 10, \tag{5.1}$$

where v_i and e_{ij} are normally distributed errors with common variances σ_v^2 and σ_e^2. We fitted model (5.1) to the sample data to obtain EBP estimates of $\boldsymbol{\beta}$ and v_i, denoted as $\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2\right)^T$ and $\hat{v}_i$, and re-parameterized REML estimates of the variance components, denoted as $\left(\hat{\sigma}_v^{2\text{reRE}}, \hat{\sigma}_e^{2\text{reRE}}\right)$. The EBLUP estimates of the model fixed effects are $\hat{\beta}_0 = 58.5$, $\hat{\beta}_1 = 0.316$ and $\hat{\beta}_2 = -0.150$, whereas the reREML estimates of the variance components are $\hat{\sigma}_v^{2\text{reRE}} = 135.6$ and $\hat{\sigma}_e^{2\text{reRE}} = 155.9$. The estimated δ is 0.869 which is close to 1. For each unit in the sample, we replicated the vector $\mathbf{x}_{ij} = \left(1, x_{1ij}, x_{2ij}\right)^T$ several times equal to $\left[d_{ij}\right]$, the closest integer to the sampling weight $d_{ij} = N_i / n_i$. Thus, we obtained a pseudo-population of x-values, denoted as $\mathbf{x}_{ij}^{\text{ps}} = \left(1, x_{1ij}^{\text{ps}}, x_{2ij}^{\text{ps}}\right)^T$, with county population size equal to $N_i^{\text{ps}} = n_i\left[N_i / n_i\right]$. The y-values of our pseudo-population, denoted as y_{ij}^{ps}, are defined as: $y_{ij}^{\text{ps}} = y_{ij}$ for $j \in s_i$, and $y_{ij}^{\text{ps}} = \hat{\beta}_0 + x_{1ij}^{\text{ps}}\hat{\beta}_1 + x_{2ij}^{\text{ps}}\hat{\beta}_2 + \hat{v}_i + e_{ij}^{\text{ps}}$ for $j \in r_i^{\text{ps}}$, where $e_{ij}^{\text{ps}} \sim N\left(0, \hat{\sigma}_e^{2\text{reRE}}\right)$ and r_i^{ps} is composed of the $N_i^{\text{ps}} - n_i$ non-observed units in the i^{th} small area. Prasad and Rao (1990) used a similar procedure to generate a pseudo-population with a larger number of counties than the data set provided by Battese et al. (1988). Their pseudo population composed of twenty counties was obtained in two steps: first, the values of the auxiliary variables associated with the original data set were duplicated; then, the values of the response variable were computed from the model, by using the duplicated x-values and the estimates of the model parameters.

Let $\bar{Y}_i = \left(N_i^{\text{ps}}\right)^{-1}\sum_{j=1}^{N_i^{\text{ps}}} y_{ij}^{\text{ps}}$ and $Y = \sum_{i=1}^{10} N_i^{\text{ps}}\bar{Y}_i$ be respectively the mean of the i^{th} small area and the total of the pseudo-population. At the population level we estimate Y by the GREG estimator $\hat{Y}^{\text{GREG}}$ based on weights given by (3.2) where the vector $\mathbf{x}_{ij}^{\text{ps}*}$ is the two-dimensional vector $\mathbf{x}_{ij}^{\text{ps}*} = \left(1, x_{1ij}^{\text{ps}}\right)^T$. It follows that $\mathbf{x}_{ij}^{\text{ps}} \not\subset \mathbf{x}_{ij}^{\text{ps}*}$ given that $\mathbf{x}_{ij}^{\text{ps}} = \left(1, x_{1ij}^{\text{ps}}, x_{2ij}^{\text{ps}}\right)^T$ and $\mathbf{x}_{ij}^{\text{ps}*} = \left(1, x_{1ij}^{\text{ps}}\right)^T$.

From the pseudo-population $\left(y_{ij}^{\text{ps}}, \mathbf{x}_{ij}^{\text{ps}}\right)$, $j = 1, \ldots, N_i^{\text{ps}}$; $i = 1, \ldots, 10$, we drew $G = 30{,}000$ stratified simple random samples without replacement of size n_i, and treating each county as a stratum. These sample sizes were equal to those of the original data set. We used the design relative bias (RB) and mean squared error (RRMSE) to evaluate the performance of six estimators: two non benchmarked estimators, $\hat{\bar{Y}}_i^{\text{EBLUP}}$ and $\hat{\bar{Y}}_i^{\text{YR}}$, and four benchmarked estimators, $\hat{\bar{Y}}_{ib}^{\text{EBRat}}$, $\hat{\bar{Y}}_{ib}^{\text{YRat}}$, $\hat{\bar{Y}}_{ib}^{\text{REBLUP}}$ and $\hat{\bar{Y}}_{ib}^{\text{RYR}}$, that can be computed in the case $\mathbf{x}_{ij}^{\text{ps}} \not\subset \mathbf{x}_{ij}^{\text{ps}*}$. Let $\hat{\bar{Y}}_i$ be a generic estimator of the i^{th} small area mean $\bar{Y}_i$, and $\hat{\bar{Y}}_i^{(g)}$ its value associated with the g^{th} sample, for $g = 1, \ldots, G$. Its RB and RRMSE values are given by

$$\text{RB}_i = \frac{1}{G}\sum_{g=1}^{G}\frac{\hat{\bar{Y}}_i^{(g)}}{\bar{Y}_i} - 1 \quad \text{and} \quad \text{RRMSE}_i = \sqrt{\frac{1}{G}\sum_{g=1}^{G}\left(\frac{\hat{\bar{Y}}_i^{(g)}}{\bar{Y}_i} - 1\right)^2}.$$

Table 5.1 reports on the design RB and RRMSE of the six estimators of $\bar{Y}_i$ for the ten counties of the pseudo population. From this example, we see that the RBs and RRMSEs are quite similar across all estimators and sample sizes. This follows because the model that generated the population data is correct, whereas both the small area model and the GREG estimator have in common the auxiliary variable equal to the number of pixels classified as corn.

Table 5.1
RB (%) and RRMSE (%): the benchmark to $\hat{Y}^{\text{GREG}}\left(\mathbf{x}_{ij}^{\text{ps}} \not\subset \mathbf{x}_{ij}^{\text{ps*}}\right)$

County	n_i	Measure	$\hat{\bar{Y}}_i^{\text{EBLUP}}$	$\hat{\bar{Y}}_i^{\text{YR}}$	$\hat{\bar{Y}}_{ib}^{\text{EBRat}}$	$\hat{\bar{Y}}_{ib}^{\text{YRat}}$	$\hat{\bar{Y}}_{ib}^{\text{REBLUP}}$	$\hat{\bar{Y}}_{ib}^{\text{RYR}}$
Cerro Hamilton Worth	3	RB	1.6	1.4	1.3	1.3	1.0	1.2
		RRMSE	5.2	5.4	5.3	5.4	5.6	5.4
Humboldt	2	RB	2.0	1.9	1.7	1.8	1.8	1.8
		RRMSE	4.5	4.5	4.5	4.5	4.4	4.5
Franklin	3	RB	-3.3	-3.4	-3.5	-3.5	-3.5	-3.5
		RRMSE	5.2	5.4	5.5	5.5	5.4	5.4
Pocahontas	3	RB	-3.1	-3.4	-3.4	-3.5	-3.3	-3.5
		RRMSE	6.2	6.5	6.4	6.6	6.4	6.6
Winnebago	3	RB	2.6	2.3	2.3	2.2	2.3	2.2
		RRMSE	5.4	5.3	5.3	5.3	5.3	5.2
Wright	3	RB	-0.4	-0.6	-0.7	-0.7	-0.6	-0.6
		RRMSE	3.7	3.8	3.9	3.9	3.8	3.9
Webster	4	RB	-2.6	-2.9	-2.9	-3.0	-2.8	-2.9
		RRMSE	5.2	5.4	5.5	5.5	5.4	5.5
Hancock	5	RB	0.9	0.7	0.6	0.6	0.8	0.7
		RRMSE	4.2	4.1	4.2	4.2	4.2	4.2
Kossuth	5	RB	3.5	3.3	3.2	3.2	3.2	3.2
		RRMSE	5.9	5.8	5.8	5.8	5.8	5.8
Hardin	5	RB	-1.5	-1.7	-1.8	-1.8	-1.7	-1.8
		RRMSE	4.2	4.3	4.4	4.5	4.3	4.4

6. Conclusion

In general, the sum of model-based small area estimates is not equal to a direct estimate obtained across the union of these small areas. The weight that is associated with the direct estimator can be the sampling weight or one obtained as a result of using the GREG estimator. The auxiliary data that are used to obtain the GREG and the unit-level small area estimates may not necessarily coincide. In this paper, we have suggested several benchmarking procedures for two well-known small area estimators (EBLUP and YR) that are based on the unit level model. We considered the case when the sampling rates are not negligible, and that the sample design is ignorable. In the event that it is deemed that the sample design is not ignorable for some of the survey items, the auxiliary data vector $\mathbf{x}_{ij}$ in model (2.2) could be augmented by including an additional variable g_{ij} specified function of the survey weights to offset the potential bias of the EBLUP or YR estimators. Verret et al. (2015) proposed a number of choices for g_{ij} that included the survey weight w_{ij}. In the case of the EBLUP estimator, benchmarking is achieved by adding the variable $q_{ij} = w_{ij}^{\text{GREG}} - 1$. Since q_{ij} should be highly correlated to w_{ij}, the suggested procedure for benchmarking EBLUP should provide good protection against possible non ignorable sampling. The simulations in Verret et al. (2015) illustrated that the YR procedure, on its own, provides good protection as well against possible non ignorable sampling. Their simulation also showed that further protection can be obtained by their setting g_{ij} equal to $n_i w_{ij}$.

We extended the benchmarking procedures in Stefan and Hidiroglou (2020) to the case of non-negligible sampling rates within each small area. These procedures are based on estimators that were initially developed by Battese et al. (1988) (EBLUP estimator), and You and Rao (2002) (YR estimator)

when the sampling rates within each small area are negligible. Ugarte et al. (2009) proposed a different benchmarked estimator which is a restricted EBLUP estimator. We extended the procedure in Ugarte et al. (2009) to obtain a benchmarked estimator that incorporates the survey weights, and that is essentially a restricted YR estimator. We also considered two benchmarked estimators based on simple ratio adjustments applied on the EBLUP and YR estimators respectively. We carried out a simulation study to compare the properties of these six benchmarked estimators.

If the auxiliary data used to estimate the small area means are the same as those used in the GREG, and if the model is correct, the restricted procedure in Ugarte et al. (2009) and the ratio adjusted EBLUP estimator will have the smaller $\overline{\text{ARB}}$'s and $\overline{\text{RRMSE}}$'s. On the other hand, if the model is incorrect and the auxiliary data are the same ones, the YR estimator based on Stefan and Hidiroglou (2020) procedure, adapted to non-negligible sampling rates, has the smallest $\overline{\text{ARB}}$'s, whereas the restricted YR estimator has the smallest $\overline{\text{RRMSE}}$'s. On the other hand, if the auxiliary data used to estimate the small area means are not the same as those used in the GREG, we come to the following conclusions. The restricted EBLUP and the ratio adjusted EBLUP estimators are the benchmarked estimators that have the smallest $\overline{\text{ARB}}$'s and $\overline{\text{RRMSE}}$'s if the model is correct. If the model is not correct, the restricted YR estimator is the preferred choice both in terms of $\overline{\text{ARB}}$ and $\overline{\text{RRMSE}}$.

Benchmarking should be based on the EBLUP procedure if the linear mixed effects model is appropriate. If the linear model and the benchmark (the GREG estimator) have in common a large amount of auxiliary information, the benchmarked estimators are similar to their non benchmarked versions, otherwise the loss of efficiency due to benchmarking may be important. If the model is not correct, the YR procedure should be used to achieve benchmarking. In this case, benchmarking may bring about important gains in terms of $\overline{\text{ARB}}$ and $\overline{\text{RRMSE}}$, especially if the small area model and the GREG estimator share a small number of auxiliary variables. The finite populations associated with incorrect modeling were generated based on model (4.2), with mean function incorrectly specified. However, there are many ways in which a model may be wrong, and the conclusions associated with these cases may be different.

Acknowledgements

We would like to thank the two anonymous referees and the associate editor for their constructive suggestions.

Appendix A

Proof of Result 1. The EBLUP estimators $\hat{\boldsymbol{\beta}}_a = \left(\left(\hat{\boldsymbol{\beta}}_{1a} \right)^T, \hat{\beta}_{2a} \right)^T$ and $\hat{v}_{ia}$, that are based on model (3.6), satisfy the equation

$$\sum_{i=1}^{m} \sum_{j \in s_i} \begin{pmatrix} \mathbf{x}_{ij} \\ q_{ij} \end{pmatrix} \left(y_{ij} - \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}_{1a} - q_{ij} \hat{\beta}_{2a} - \hat{v}_{ia} \right) = 0. \tag{A.1}$$

Equation (A.1) has the form of equation (2.10) and corresponds to augmented model (3.6). Expanding the second equation in (A.1), we obtain that

$$\sum_{i=1}^{m}\sum_{j\in s_i} q_{ij}\mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}_{1a} + \sum_{i=1}^{m}\sum_{j\in s_i} q_{ij}^2\hat{\beta}_{2a} + \sum_{i=1}^{m}\sum_{j\in s_i} q_{ij}\hat{v}_{ia} = \sum_{i=1}^{m}\sum_{j\in s_i} q_{ij}y_{ij}. \tag{A.2}$$

The variable q_{ij} is defined as $q_{ij} = w_{ij}^{\mathrm{GREG}} - 1$. The right-hand side of (A.2) is

$$\sum_{i=1}^{m}\sum_{j\in s_i} q_{ij}y_{ij} = \sum_{i=1}^{m}\sum_{j\in s_i}\left(w_{ij}^{\mathrm{GREG}} - 1\right) y_{ij} = \hat{Y}^{\mathrm{GREG}} - \sum_{i=1}^{m}\sum_{j\in s_i} y_{ij}. \tag{A.3}$$

The sums that appear on the left-hand side of (A.2) are given by

$$\sum_{i=1}^{m}\sum_{j\in s_i} q_{ij}\mathbf{x}_{ij}^T = \sum_{i=1}^{m}\sum_{j\in s_i}\left(w_{ij}^{\mathrm{GREG}} - 1\right)\mathbf{x}_{ij}^T = \left(\sum_{i=1}^{m}\mathbf{X}_i - \sum_{j\in s_i}\mathbf{x}_{ij}\right)^T = \left(\sum_{i=1}^{m}\mathbf{x}_{ir}\right)^T, \tag{A.4}$$

$$\sum_{i=1}^{m}\sum_{j\in s_i} q_{ij}^2 = \sum_{i=1}^{m} q_{iw}, \tag{A.5}$$

$$\sum_{i=1}^{m}\sum_{j\in s_i} q_{ij} = \sum_{i=1}^{m}\sum_{j\in s_i}\left(w_{ij}^{\mathrm{GREG}} - 1\right) = \sum_{i=1}^{m}\left(\hat{N}_i^{\mathrm{GREG}} - n_i\right). \tag{A.6}$$

In establishing that last equality of (A.4), we used that $\mathbf{x}_{ij} \subseteq \mathbf{x}_{ij}^*$ and that weights w_{ij}^{GREG} satisfy equation (3.3). Result 1 follows by replacing (A.3), (A.4), (A.5) and (A.6) into (A.2).

Proof of Result 2. The survey-weighted estimating equations that defines $\hat{\boldsymbol{\beta}}^{\mathrm{YR}}$ and $\hat{\mathbf{v}}^{\mathrm{YR}}$ are given by (2.12) constructed with the weights $w_{ij}^{\mathrm{GREG}} - 1$:

$$\sum_{i=1}^{m}\sum_{j\in s_i}\left(w_{ij}^{\mathrm{GREG}} - 1\right)\mathbf{x}_{ij}\left(y_{ij} - \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}^{\mathrm{YR}} - \hat{v}_i^{\mathrm{YR}}\right) = 0.$$

Since the first term of $\mathbf{x}_{ij}$ is one (representing an intercept), it follows that

$$\sum_{i=1}^{m}\sum_{j\in s_i}\left(w_{ij}^{\mathrm{GREG}} - 1\right)\left(y_{ij} - \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}^{\mathrm{YR}} - \hat{v}_i^{\mathrm{YR}}\right) = 0. \tag{A.7}$$

The terms in (A.7) are given by:

$$\sum_{i=1}^{m}\sum_{j\in s_i}\left(w_{ij}^{\mathrm{GREG}} - 1\right) y_{ij} = \hat{Y}^{\mathrm{GREG}} - \sum_{i=1}^{m}\sum_{j\in s_i} y_{ij}, \tag{A.8}$$

$$\sum_{i=1}^{m}\sum_{j\in s_i}\left(w_{ij}^{\mathrm{GREG}} - 1\right)\mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}^{\mathrm{YR}} = \left(\sum_{i=1}^{m}\mathbf{X}_i - \sum_{j\in s_i}\mathbf{x}_{ij}\right)^T\hat{\boldsymbol{\beta}}^{\mathrm{YR}} = \left(\sum_{i=1}^{m}\mathbf{x}_{ir}\right)^T\hat{\boldsymbol{\beta}}^{\mathrm{YR}}, \tag{A.9}$$

and

$$\sum_{i=1}^{m}\sum_{j\in s_i}\left(w_{ij}^{\mathrm{GREG}}-1\right)\hat{v}_i^{\mathrm{YR}}=\sum_{i=1}^{m}\left(\hat{N}_i^{\mathrm{GREG}}-n_i\right)\hat{v}_i^{\mathrm{YR}}. \tag{A.10}$$

Plugging (A.8), (A.9) and (A.10) into (A.7) leads to

$$\sum_{i=1}^{m}\sum_{j\in s_i}y_{ij}+\left(\sum_{i=1}^{m}\mathbf{x}_{ri}\right)^{T}\hat{\boldsymbol{\beta}}^{\mathrm{YR}}+\sum_{i=1}^{m}\left(\hat{N}_i^{\mathrm{GREG}}-n_i\right)\hat{v}_i^{\mathrm{YR}}=\hat{Y}^{\mathrm{GREG}}. \tag{A.11}$$

Equation (A.11) proves Result 2.

Appendix B

Re-parameterized REML estimation of variance components

Let $\boldsymbol{\delta}=(\delta_1,\,\delta_2)$ be the vector of variance components, where $\delta_1=\sigma_v^2$ and $\delta_2=\sigma_e^2$. We define the vector $\boldsymbol{\alpha}=(\alpha_1,\,\alpha_2)$ such that $\sigma_v^2=e^{\alpha_1}$ and $\sigma_e^2=e^{\alpha_2}$. The restricted maximum log-likelihood function, denoted as $l(\boldsymbol{\alpha})$ is

$$l(\boldsymbol{\alpha})=l(\alpha_1,\,\alpha_2)=c-\frac{1}{2}\log|\mathbf{V}|-\frac{1}{2}\log|\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}|-\frac{1}{2}\mathbf{y}^T\mathbf{P}\mathbf{y}, \tag{B.1}$$

where c is a generic constant, $\mathbf{V}=e^{\alpha_1}\mathbf{Z}\mathbf{Z}^T+e^{\alpha_2}\mathbf{I}_n$ and $\mathbf{P}=\mathbf{V}^{-1}-\mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}$. Notice that $\mathbf{P}\mathbf{X}=\mathbf{0}$. The solution to the maximization of $l(\boldsymbol{\alpha})$ is obtained iteratively using the Fisher-scoring algorithm by updating the following equation

$$\boldsymbol{\alpha}^{(r+1)}=\boldsymbol{\alpha}^{(r)}+\mathbf{I}\left(\boldsymbol{\alpha}^{(r)}\right)^{-1}\mathbf{s}\left(\boldsymbol{\alpha}^{(r)}\right). \tag{B.2}$$

Here, $\mathbf{s}\left(\boldsymbol{\alpha}^{(r)}\right)=\left(\partial l\left(\boldsymbol{\alpha}^{(r)}\right)/\partial\alpha_1,\,\partial l\left(\boldsymbol{\alpha}^{(r)}\right)/\partial\alpha_2\right)^T$ is the vector of first-order partial derivatives of $l(\boldsymbol{\alpha})$ with respect to $\boldsymbol{\alpha}$, and $\mathbf{I}\left(\boldsymbol{\alpha}^{(r)}\right)=\left(I_{jk}\left(\boldsymbol{\alpha}^{(r)}\right)\right)_{j,k=1,2}$ is the matrix of expected second-order derivatives of $-l(\boldsymbol{\alpha})$ with respect to $\boldsymbol{\alpha}$, where $I_{jk}\left(\boldsymbol{\alpha}^{(r)}\right)=E\left(-\partial^2 l\left(\boldsymbol{\alpha}^{(r)}\right)/\partial\alpha_j\partial\alpha_k\right)$.

Under the BHF model, the first-order partial derivatives of $l(\boldsymbol{\alpha})$ are given by

$$\frac{\partial l}{\partial\alpha_j}(\boldsymbol{\alpha})=\left[-\frac{1}{2}\mathrm{tr}\left(\mathbf{P}\mathbf{V}_{(j)}\right)+\frac{1}{2}\mathbf{y}^T\mathbf{P}\mathbf{V}_{(j)}\mathbf{P}\mathbf{y}\right]e^{\alpha_j},\ j=1,\,2, \tag{B.3}$$

where $\mathbf{V}_{(1)}=\mathbf{Z}\mathbf{Z}^T$ and $\mathbf{V}_{(2)}=\mathbf{I}_n$. The expected values of the second-order partial derivatives of $l(\boldsymbol{\alpha})$ are

$$E\left(\frac{\partial^2 l}{\partial\alpha_j\partial\alpha_k}(\boldsymbol{\alpha})\right)=-\frac{1}{2}\mathrm{tr}\left(\mathbf{P}\mathbf{V}_{(j)}\mathbf{P}\mathbf{V}_{(k)}\right)e^{\alpha_j+\alpha_k},\ j,\,k=1,\,2. \tag{B.4}$$

The re-parameterized REML estimator of $\boldsymbol{\delta}$ is obtained as

$$\hat{\boldsymbol{\delta}}^{\mathrm{reRE}}=\left(\hat{\sigma}_v^{2\mathrm{reRE}},\,\hat{\sigma}_e^{2\mathrm{reRE}}\right)=\left(e^{\hat{\alpha}_1},\,e^{\hat{\alpha}_2}\right). \tag{B.5}$$

References

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

Bell, W.R., Datta, G.S. and Ghosh, M. (2013). Benchmarking small area estimators. *Biometrika*, 100, 189-202.

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

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

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

Hidiroglou, M.A., and Estevao, V.M. (2016). A comparison of small area and traditional estimators via simulation. *Statistics in Transition*, 17, 133-154.

Huang, R., and Hidiroglou, M.A. (2003). Design consistent estimators for a mixed linear model on survey data. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 1897-1904.

Moghtased-Azar, K., Tehranchi, R. and Amiri-Simkooei, A.R. (2014). An alternative method for non-negative estimation of variance components. *Journal of Geodesy*, 88, 427-439.

Nandram, B., and Sayit, H. (2011). A Bayesian analysis of small area probabilities under a constraint. *Survey Methodology*, 37, 2, 137-152. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2011002/article/11603-eng.pdf.

Pfeffermann, D., and Barnard, C. (1991). Some new estimators for small area means with applications to the assessment of farmland values. *Journal of Business and Economic Statistics*, 9, 73-83.

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

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

Särndal, C.-E., Swensson, B. and Wretman, J.H. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76, 527-537.

Stefan, M., and Hidiroglou, M.A. (2020). Benchmarked estimators for a small area mean under a one-fold nested regression model. *International Statistical Review*, (To appear).

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

Ugarte, M.D., Militino, A.F. and Goicoa, T. (2009). Benchmarked estimates in small areas using linear mixed models with restrictions. *Test*, 18, 342-364.

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

Wang, J., Fuller, W.A. and Qu, Y. (2008). Small area estimation under a restriction. *Survey Methodology*, 34, 1, 29-36. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2008001/article/10619-eng.pdf.

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

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

You, Y., Rao, J.N.K. and Hidiroglou, M.A. (2013). On the performance of self benchmarked small area estimators under the Fay-Herriot area level model. *Survey Methodology*, 39, 1, 217-229. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2013001/article/11830-eng.pdf.

Survey Methodology, June 2021
Vol. 47, No. 1, pp. 151-189
Statistics Canada, Catalogue No. 12-001-X

Estimation of domain discontinuities using Hierarchical Bayesian Fay-Herriot models

Jan A. van den Brakel and Harm-Jan Boonstra[1]

Abstract

Changes in the design of a repeated survey generally result in systematic effects in the sample estimates, which are further referred to as discontinuities. To avoid confounding real period-to-period change with the effects of a redesign, discontinuities are often quantified by conducting the old and the new design in parallel for some period of time. Sample sizes of such parallel runs are generally too small to apply direct estimators for domain discontinuities. A bivariate hierarchical Bayesian Fay-Herriot (FH) model is proposed to obtain more precise predictions for domain discontinuities and is applied to a redesign of the Dutch Crime Victimization Survey. This method is compared with a univariate FH model where the direct estimates under the regular approach are used as covariates in a FH model for the alternative approach conducted on a reduced sample size and a univariate FH model where the direct estimates for the discontinuities are modeled directly. An adjusted step forward selection procedure is proposed that minimizes the WAIC until the reduction of the WAIC is smaller than the standard error of this criteria. With this approach more parsimonious models are selected, which prevents selecting complex models that tend to overfit the data.

Key Words: Area level models; Bivariate Fay-Herriot model; Small area estimation; Survey redesign; Measurement bias; MCMC; Gibbs sampler.

1. Introduction

Official statistics produced by national statistical institutes are generally based on repeated sample surveys. Much of their value lies in their continuity, enabling developments in society and the economy to be monitored, and policy actions decided. Survey samples contain besides sampling errors different sources of non-sampling errors that have a systematic effect on the outcomes of a survey. As long as the survey process is kept constant, this bias component is not visible. This is often an argument to keep survey processes of repeated surveys unchanged as long as possible. From time to time changes in surveys are needed to improve the efficiency, reduce the survey related costs, or meet new requirements, and this is seen strongly in the use of mixed-mode surveys including web-based questionnaires in official statistics. A redesign of the survey process generally has systematic effects on the survey estimates, since the biases induced by the aforementioned non-sampling errors are changed, disturbing comparability with figures published in the past.

Systematic differences in the outcomes of a repeated survey due to redesign of the survey process are called discontinuities. To avoid the implementation of a new survey process disturbing the comparability of estimates over time, it is important to quantify these discontinuities. This avoids confounding real change in the parameters of interest with changing measurement bias due to alteration of the survey process.

1. Jan A. van den Brakel, Statistics Netherlands, Department of Statistical Methods and Maastricht University, Department of Quantitative Economics. E-mail: ja.vandenbrakel@cbs.nl; Harm-Jan Boonstra, Statistics Netherlands, Department of Statistical Methods.

Several methods to quantify discontinuities are proposed in the literature (van den Brakel, Smith and Compton, 2008). A reliable and straightforward approach is to conduct the old and new approach alongside of each other at the same time for some period of time, further referred to as a parallel run. Ideally this is based on a randomized experiment that can be embedded in the probability sample of the survey (van den Brakel, 2008). In this paper small area estimation methods for estimating domain discontinuities are proposed. We consider the situation where the regular survey, used for the production of official figures, is conducted at the full sample size and is conducted in parallel with an alternative approach. Due to budget limitations, the sample that is assigned to the alternative approach is often not sufficiently large to observe minimum detectable differences at prespecified significance and power levels using standard direct estimators, particularly for sub populations or domains.

To explain the problem addressed in this paper, some notation is introduced. Let θ_i denote the real population value of a variable of interest for domain i. Furthermore, $\hat{y}_i^r$ and $\hat{y}_i^a$ denote direct estimates of θ_i based on the regular survey and the alternative survey approach, respectively. Since the regular survey is conducted at the regular sample size, $\hat{y}_i^r$ is a reliable direct estimate for θ_i, at least for the planned domains. Due to the reduced sample size of the new survey in the parallel run $\hat{y}_i^a$, however, will be insufficiently precise. More precise domain estimates with the small sample available under the new approach can be obtained with the Fay-Herriot (FH) model (Fay and Herriot, 1979), which is defined as $\hat{y}_i^a = \mathbf{x}_i^t\beta + \nu_i + e_i^a$, with $\mathbf{x}_i$ a vector with covariates at the domain level, β the regression coefficients, ν_i the random domain effects and e_i^a the sampling error. To obtain more precise domain estimates for the alternative approach, van den Brakel, Buelens and Boonstra (2016) proposed an hierarchical Bayesian (HB) univariate FH model, where sample estimates of the regular survey are considered as potential auxiliary variables in a model selection procedure. This implies that $\hat{y}_i^r$ is used as a covariate in $\mathbf{x}_i$, besides the usual covariates that are available from registers or censuses. This results in an area level model, with measurement error (Ybarra and Lohr, 2008). The use of reliable direct estimates observed in the regular survey significantly increased the precision of the domain estimates for the alternative approach conducted at reduced sample size (van den Brakel et al., 2016).

Let $\tilde{y}_i^a$ denote the small area prediction for θ_i based on the aforementioned FH model under the small sample assigned to the alternative survey approach. In the approach followed by van den Brakel et al. (2016), point estimates for domain discontinuities are obtained as the difference between the direct estimate obtained with the regular survey and the model based domain prediction obtained under the alternative approach, i.e., $\tilde{\Delta}_i = \hat{y}_i^r - \tilde{y}_i^a$. The use of the direct estimate of the regular survey as an auxiliary variable in the small domain predictions of the alternative survey, results in strong positive correlations between both estimators, which cannot be ignored when computing the standard errors for the discontinuities. More precisely, $\mathrm{Var}(\tilde{\Delta}_i) = \mathrm{Var}(\hat{y}_i^r) + \mathrm{MSE}(\tilde{y}_i^a) - 2\mathrm{Cov}(\hat{y}_i^r, \tilde{y}_i^a)$. Since $\hat{y}_i^r$ is also used as a covariate in $\mathbf{x}_i$ in the FH model for $\tilde{y}_i^a$, $\mathrm{Cov}(\hat{y}_i^r, \tilde{y}_i^a)$ will be nonzero. To this end, two analytic approximations for the standard errors of the discontinuities are proposed. The first approach combines the design-based variance estimate of the direct estimator of the regular survey $(\mathrm{Var}(\hat{y}_i^r))$ with the posterior

variance of the HB domain predictions of the alternative survey $\left(\text{MSE}\left(\tilde{y}_i^a\right)\right)$ and a design-based estimator for the covariance between both point estimates $\left(\text{Cov}\left(\hat{y}_i^r, \tilde{y}_i^a\right)\right)$. This approach is unstable in the sense that even negative variance estimates occur in the case of strong positive covariance estimates. A related issue is that design-based and model-based variance approximations are combined in one uncertainty measure for the discontinuities. Therefore a second analytic approximation was proposed, where a design-based estimator for the variance of the HB domain predictions $\left(\text{MSE}\left(\tilde{y}_i^a\right)\right)$ is derived and combined with the design-based variance for the direct estimator for the regular survey and the design-based covariance between both point estimates.

Several references to design-based mean squared error estimation in small area estimation can be found in the literature. Gonzalez and Waksberg (1973) introduced the concept of an average design-based mean squared error of a set of synthetic estimators and proposed an estimator that, however, can be unstable and take negative values. Marker (1995) proposed a more stable but biased estimator for the design-based mean squared error for small area estimates, which can also take negative values. Lahiri and Pramanik (2019) proposed a design-based estimator that cannot take negative values, following the concepts of an average design-based mean squared error, originally introduced by Gonzalez and Waksberg (1973). Rivest and Belmonte (2000) proposed an estimator for the mean squared error that measures the uncertainty with respect to the sampling design conditional on the random effects of the model and assuming normality of the sampling model. Rao, Rubin-Bleuer and Estevao (2018) and Pfeffermann and Ben-Hur (2018) also propose a model for the design-based mean squared error in small area estimators. Rao et al. (2018) estimate the model parameters through restricted maximum likelihood while Pfeffermann and Ben-Hur (2018) applies a bootstrap method.

The complications with variance estimation of domain discontinuities under a univariate FH model can also be circumvented by setting up a full Bayesian framework for the analysis of the domain discontinuities. Two approaches are proposed in this paper. The first approach is a bivariate FH model to model the direct estimates under the regular and alternative approach simultaneously, i.e., a bivariate area level model for the vector $\left(\hat{y}_i^r, \hat{y}_i^a\right)^t$. The random component of this model accounts for the correlation between the domain parameters under the regular and alternative approach. The precision of the estimated discontinuities is improved by increasing the effective sample size within the domains by means of cross-sectional correlations. In addition, a positive correlation between the random domain effects further decreases the standard error of the estimated discontinuities. The second approach uses a univariate FH model for the direct estimates of the discontinuities, i.e., a univariate FH model for $\hat{\Delta}_i = \hat{y}_i^r - \hat{y}_i^a$. This method is considered as a less complex alternative for the bivariate FH model. It is, however, anticipated that it is harder to construct good prediction models, since the available covariates from registers might be good predictors for the target variables of the sample survey but probably not for systematic differences between the differences of two estimates for the same variable obtained with different survey processes.

The univariate FH model proposed by van den Brakel et al. (2016) was applied to estimate domain discontinuities in five key target variable of the Dutch Crime Victimization Survey (CVS) using data

obtained in a parallel run where the regular survey is conducted at the regular sample size and the alternative survey at a sample size that is about one fourth of the regular sample size. In this paper the bivariate FH model and the univariate FH model for the domain discontinuities are also applied to the same redesign of the CVS. The results are compared with the univariate FH model proposed in van den Brakel et al. (2016).

Model selection in this paper is based on a step forward selection procedure that minimizes the WAIC criteria (Watanabe, 2010, 2013). To avoid selecting over-parameterized models, it is proposed to add covariates in a step forward selection procedure only if they decrease the WAIC by more than the standard error of the WAIC. This prevents selection of several covariates that only marginally improves the WAIC, resulting in models that tend to overfit the data.

The FH model (Fay and Herriot, 1979) is frequently applied in the context of small area estimation (Rao and Molina, 2015). FH models are particularly appropriate if auxiliary information is available at the domain level. Datta, Ghosh, Nangia and Natarjan (1996) employed a multivariate FH model fitted in an HB framework to estimate median income. Multivariate FH models fitted in a frequentist framework are considered in Gonzales-Manteiga, Lombardia, Molina, Morales and Santamaria (2008); Benavent and Morales (2016). Several authors provided time-series FH models to use sample information from previous editions of a survey as a form of small area estimation (Rao and Yu, 1994; Datta, Lahiri, Maiti and Lu, 1999; You and Rao, 2000; Estaban, Morales, Perez and Santamaria, 2012; Marhuenda, Molina and Morales, 2013). Pfeffermann and Burck (1990); Pfeffermann and Tiller (2006); van den Brakel and Krieg (2016); Bollineni-Balabay, van den Brakel, Palm and Boonstra (2017) are some examples of FH time-series models casted in a state-space framework. Boonstra and van den Brakel (2019) discuss how FH time series models can be expressed either in a state space frame work and fitted with the Kalman filter or alternatively expressed as time series multilevel models in an hierarchical Bayesian framework, and estimated using a Gibbs sampler.

The paper is structured as follows. In Section 2 the Crime Vicitimization Survey, the redesign and the set up of the parallel run are described. The bivariate FH model is explained in Section 3, including the HB framework and the model selection and evaluation approach. Results are presented in Section 4. The paper ends with a discussion in Section 5.

2. The crime victimization survey

The Dutch crime victimization survey (CVS) is a long-standing survey conducted by Statistics Netherlands at an annual frequency with the purpose to publish reliable figures about crime rates, safety feelings, and satisfaction about police performance in the Netherlands. The CVS is designed to provide reliable figures at the national level and at the level of police districts, which is a subdivision of the Netherlands in 25 regions. The CVS is based on a stratified simple random sampling design for people aged 15 years or older residing in the Netherlands. Strata are formed by police regions to control the

precision of these planned domain estimates. The sampling frame is based on the Dutch government's register of all residents in the Netherlands, called Municipal Basis Administration. The yearly sample of the regular CVS is designed such that about 19,000 respondents are observed. The sample is equally divided over the strata, such that about 760 observations are obtained in each stratum. The general regression (GREG) estimator (Särndal, Swensson and Wretman, 1992) is used to estimate population parameters at the national level and for police districts.

The CVS has been redesigned in 2008. The data collection changed from a mixed-mode design via computer-assisted personal interviewing (CAPI) and computer-assisted telephone interviewing (CATI) to a sequential mixed-mode design that starts with web interviewing (WI) and is followed up for nonrespondents with CAPI and CATI. In addition the questionnaire is changed to improve the wording as well as the order of the questions. To quantify discontinuities induced by this redesign, the regular survey used for official publication purposes was conducted in parallel with the alternative survey approach with a sample size of about 6,000 respondents. In this application, the regular approach was based on the new survey design using WI, CATI and CAPI and the alternative approach was based on the old design using CAPI and CATI data collection only. The sample design for the parallel run is based on stratified simple random sampling where police districts are the strata, using proportional allocation. This results in a sample design that is optimal to estimate figures at the national level but suboptimal for domain estimation.

This survey reports on many different outcome variables. In the present study five key survey variables are considered, see Table 2.1. Estimates for these variables at the national level under the regular and alternative survey are specified in Table 2.2. The sample size allocated to the alternative approach is sufficiently large to estimate discontinuites at the national level using the GREG estimator but insufficient to estimate discontinuities at the domain level of the 25 police districts. The direct estimates for the discontinuities at the national level are indeed significantly different from zero, contrary to the unweighted averages of the direct domain estimates and their standard errors. To obtain more precise predictions for domain discontinuities a model-based small area estimation method based on area level models (Fay and Herriot, 1979) is proposed in the next section.

Table 2.1
Five key CVS survey variables considered in the present study

Variable	Description
nuisance	Perceived nuisance in the neighborhood on a ten point scale; this includes nuisance by drunk people, neigbours, or groups of youngsters, harassment, and drug related problems.
unsafe	Percentage of people feeling unsafe at times.
propvict	Percentage of people saying to have been victim to property crime in the last 12 months.
offtot	Total number of offenses per 100 people.
satispol	Percentage of people satisfied with police at their last contact (if contact in last 12 months).

Table 2.2
GREG estimates for the regular and alternative survey approach averaged over districts and national level. Standard errors between brackets

Variable	Average over 25 police districts						National level					
	regular		alternative		$\hat{\Delta}$		regular		alternative		$\hat{\Delta}$	
offtot	42.29	(4.73)	33.28	(5.73)	9.01	(7.69)	43.79	(1.07)	34.09	(1.04)	9.70	(1.49)
unsafe	24.38	(2.03)	19.86	(2.87)	4.52	(3.57)	25.07	(0.44)	20.48	(0.52)	4.59	(0.68)
nuisance	1.61	(0.11)	1.28	(0.13)	0.33	(0.17)	1.67	(0.02)	1.34	(0.02)	0.33	(0.03)
satispol	60.61	(4.23)	55.58	(6.88)	5.04	(8.21)	59.88	(0.92)	55.10	(1.25)	4.78	(1.55)
propvict	12.55	(1.60)	9.78	(2.19)	2.78	(2.77)	13.02	(0.36)	10.32	(0.39)	2.70	(0.53)

3. Methods

3.1 Small area estimation for domain discontinuities

Testing hypotheses about differences between estimates of a finite population parameter observed under different survey processes implies the existence of measurement errors. Therefore a measurement error model is required to explain systematic differences between survey estimates for the same population parameter observed under two different survey approaches. Let θ_i denote the population parameter of domain $i = 1, \ldots, m$. Let y_i^r and y_i^a denote the observed value for θ_i in the case of a complete enumeration under the regular approach and alternative approach, respectively. Direct estimates for y_i^r and y_i^a are obtained with the GREG estimator based on the samples assigned to the regular and alternative survey and are denoted as $\hat{y}_i^r$ and $\hat{y}_i^a$ respectively.

The relation between the observed values under a complete enumeration and the real population parameter is: $y_i^q = \theta_i + \lambda_i^q, i = 1, \ldots, m, q = r, a,$ with λ_i^q the real measurement bias if θ_i is measured with survey approach q. Without any external information, it is not possible to estimate λ_i^q. From the sample data only the relative bias, say $\Delta_i = y_i^r - y_i^a = \lambda_i^r - \lambda_i^a$ is identifiable. Direct estimates for these discontinuities are obtained from the survey data as the contrast between the GREG estimates, i.e., $\hat{\Delta}_i = \hat{y}_i^r - \hat{y}_i^a$.

In the case of the Dutch CVS the sample size of the regular survey is large enough to obtain sufficiently precise direct estimates for the planned domains, since the sample is designed to publish official statistics for these domains. The sample assigned to the alternative survey for the parallel run has only a size of one third of the regular sample size, which is insufficient to obtain precise direct estimates for the planned domains. In an earlier paper (van den Brakel et al., 2016) univariate FH models were developed to obtain more precise predictions for the domain parameters observed with the small sample size assigned to the alternative survey approach using auxiliary variables derived from three different sources. The first source contains demographic variables derived from the Municipal Basis Administration (MBA), which is an administration of all people residing in the Netherlands. The second source contains related variables available in the Police Register of Reported Offences (PRRO). The third source, which is unique in the case of a parallel run, contains direct estimates for the same variables observed under the

regular survey, which are sufficiently precise at least for the planned domains like police districts. The direct estimates from the regular survey are often selected as auxiliary variables for these univariate FH models. This comes not as a surprise since these are survey estimates for the same population parameters. Although measured with a different survey process, strong positive correlations can be expected. Strong improvements of the precision of small domain prediction are indeed found if the set of potential auxiliary variables, i.e., from MBA and PRRO, is extended with the direct estimates from the regular CVS.

In this application the sampling error in the auxiliary variables that come from the regular CVS can be ignored in the FH model, since the sample size and therefore the sampling error for these domains is more or less equal for the domains (Ybarra and Lohr, 2008). This implies that the variance component of the random domain effects is inflated with the sampling error of the auxiliary variables, which is fine as long as the sampling error does not differ between domains. In most applications this is not the case and the methods proposed by Ybarra and Lohr (2008) should be used to account for sampling error in the auxiliary variables.

FH multilevel models can be fitted under a frequentist approach using EBLUP or under an HB approach (Rao and Molina, 2015). In van den Brakel et al. (2016) the HB approach is preferred over the EBLUP, since the strong auxiliary information in the fixed effect part of the model often results in zero estimates for the variance component of the random domain effects, giving too much weight to the synthetic regression part and too little weight to the direct estimates in the EBLUP, (Bell, 1999; Rao and Molina, 2015). This problem can also be overcome with adjusted maximum likelihood estimation, see e.g., Li and Lahiri (2010) and Hirose and Lahiri (2018).

Let $\tilde{y}_i^a$ denote the HB prediction for domain i under the alternative approach. Now domain discontinuites are obtained by $\tilde{\Delta}_i = \hat{y}_i^r - \tilde{y}_i^a$. Using direct estimates of the regular survey as auxiliary variables in the fixed part of the FH model for the alternative survey considerably increases the complexity of the variance estimation for the discontinuities. The variance of $\tilde{\Delta}_i$ can be expressed as $\text{Var}(\tilde{\Delta}_i) = \text{Var}(\hat{y}_i^r) + \text{MSE}(\tilde{y}_i^a) - 2\text{cov}(\hat{y}_i^r, \tilde{y}_i^a)$. The use of $\hat{y}_i^r$ or related sample estimates as auxiliary variables to predict $\tilde{y}_i^a$, results in non-zero values for $\text{cov}(\hat{y}_i^r, \tilde{y}_i^a)$ that cannot be ignored. To approximate $\text{Var}(\tilde{\Delta}_i)$, van den Brakel et al. (2016) proposed an approximately design-unbiased estimator for $\text{cov}(\hat{y}_i^r, \tilde{y}_i^a)$ and $\text{Var}(\hat{y}_i^r)$, while the $\text{MSE}(\tilde{y}_i^a)$ is approximated with the posterior variance of the HB domain predictions. A major disadvantage of this approach is that model-based and design-based uncertainty measures are intertwined. On the one hand, the MSE's for $\tilde{y}_i^a$ are approximated with their posterior variances. On the other hand, the covariances between $\hat{y}_i^r$ and $\tilde{y}_i^a$ are approximated from a design-based perspective. Consequently, naive application of this approach may give negative variance estimates for the discontinuities. This drawback has been solved using a design-based approximation for the $\text{MSE}(\tilde{y}_i^a)$, resulting in a fully design-based approximation for the uncertainty of the estimated domain discontinuities.

In this paper a full HB framework for estimating domain discontinuities is proposed as an alternative by developing a bivariate FH model for the domain parameters observed under both the regular and

alternative approach. The advantage of this approach is that it improves the precision of both the direct estimates of the regular and alternative domain estimates by borrowing strength from other domains and both surveys. Negative variance estimates for the estimated domain discontinuities are precluded by definition under this multivariate HB framework. This method is compared with a simple alternative, namely a univariate FH model for the direct estimates of the discontinuities. As mentioned in the introduction, it is anticipated that it might be hard to find covariates in the available registers that explain the discontinuities. An advantage of both models is that they avoid the complications of accounting for sampling error in the auxiliary variables, which is necessary if the survey estimates of the regular survey are used as covariates in univariate FH models and the sampling error differs between domains.

3.2 Bivariate Fay-Herriot model

A bivariate version of the FH model (Fay and Herriot, 1979) starts with a measurement model for the two GREG estimates observed in each domain:

$$\hat{\mathbf{y}}_i = \mathbf{y}_i + \mathbf{e}_i,\ i = 1, \ldots, m, \tag{3.1}$$

with $\mathbf{y}_i = \left(y_i^r, y_i^a\right)^t$, $\hat{\mathbf{y}}_i$ a vector containing the GREG estimates of $\mathbf{y}_i$ and $\mathbf{e}_i = \left(e_i^r, e_i^a\right)^t$ a vector with the sampling errors of $\hat{\mathbf{y}}_i$ for which it is assumed that

$$\mathbf{e}_i \overset{\text{ind}}{\sim} \mathcal{N}\left(\mathbf{0}_2, \boldsymbol{\Psi}_i\right),\ i = 1, \ldots, m. \tag{3.2}$$

Here $\mathbf{0}_2$ is a 2 dimensional column vector with each element equal to zero. Since the sample for the regular and alternative survey are drawn independently, it is assumed that $\boldsymbol{\Psi}_i = \text{Diag}\left(\psi_i^r, \psi_i^a\right)$ where ψ_i^q is the design variance of $\hat{y}_i^q$. It is also assumed that these design variances are known although they are replaced by their estimates in practice. The true domain parameters are modelled with a multilevel model. For the fixed effects it is assumed that the regular and alternative approach share the same covariates. In the most general case the regression coefficients for the fixed part are different for both variables i.e., $y_i^q = \mathbf{x}_i^t \boldsymbol{\beta}^q + v_i^q$, with $\mathbf{x}_i = \left(x_{i1}, \ldots, x_{ip}\right)^t$ a p-vector with covariates of domain i, $\boldsymbol{\beta}^q$ a p-vector of regression coefficients, which are equal over the domains but might be different between the two survey approaches. It is assumed that $x_{i1} = 1$ corresponds to the intercept. Furthermore v_i^q are random domain effects. This gives rise to the following bivariate multilevel model for the two domain parameters:

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{v}_i,\ i = 1, \ldots, m, \tag{3.3}$$

where $\mathbf{X}_i = \mathbf{I}_2 \otimes \mathbf{x}_i^t$, $\boldsymbol{\beta} = \left(\boldsymbol{\beta}^{r^t}, \boldsymbol{\beta}^{a^t}\right)^t$, $\mathbf{I}_2$ a 2 dimensional identity matrix, and $\mathbf{v}_i = \left(v_i^r, v_i^a\right)^t$. For the random domain effects it is assumed that

$$\mathbf{v}_i \overset{\text{IID}}{\sim} \mathcal{N}\left(\mathbf{0}_2, \boldsymbol{\Sigma}\right),\ i = 1, \ldots, m, \tag{3.4}$$

with $\boldsymbol{\Sigma}$ a general 2×2 covariance matrix for the random domain effects. Inserting (3.3) into (3.1) gives:

$$\hat{\mathbf{y}}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{v}_i + \mathbf{e}_i,\ i = 1, \ldots, m, \tag{3.5}$$

with model assumptions (3.2) and (3.4).

Since the number of domains in this application is small, it is important to select parsimonious models. One way to reduce model complexity is to assume that the regression coefficients are equal for both survey approaches. In this case a dummy indicator, say δ_i, is introduced, which is equal to zero for the regular survey and equal to one for the alternative survey. In this case $x_{i1} = 1$ corresponds to the overall intercept and $x_{i2} = \delta_i$ is the indicator whose coefficient measures the differences between intercepts of the variables observed under both surveys. So $y_i^q = \mathbf{x}_i^t\boldsymbol{\beta} + \nu_i^q$ and in (3.3) $\mathbf{X}_i = \mathbf{x}_i^t$, and $\boldsymbol{\beta}$ a vector with the corresponding regression coefficients. As a result, two versions for the fixed effects are considered:

- `FE_uq`: A fixed effect model where the regular and alternative approach share the same covariates, but have different regression coefficients. In this case, domain discontinuities are given by

$$\Delta_i = \sum_{j=1}^{p} x_{i,j}\left(\beta_j^r - \beta_j^a\right) + \left(\nu_i^r - \nu_i^a\right). \tag{3.6}$$

- `FE_eq`: A more parsimonious version for the fixed effect component by assuming that the regression coefficients are equal for the regular and alternative approach. In this case domain discontinuities are given by

$$\Delta_i = -\beta_2 + \left(\nu_i^r - \nu_i^a\right), \tag{3.7}$$

  with β_2 the regression coefficient for $x_{i2} = \delta_i$.

The following covariance structures for the random domain effects are considered:

- `RE_f`: A full covariance matrix $\boldsymbol{\Sigma}$ for the random domain effects. Positive correlation between the random domain effects will further increase the precision of the estimates for the domain discontinuities since domain estimates borrow strength not only from different domains but also across the two surveys.
- `RE_d`: A diagonal covariance matrix with separate variances for the regular and alternative approach, i.e.: $\boldsymbol{\Sigma} = \text{Diag}\left(\sigma_r^2, \sigma_a^2\right)$. This covariance structure in combination with model `FE_uq` comes down to applying a univariate FH model to both surveys separately. In this case models only use sample information from other domains within the same survey but not across the two surveys to improve the precision of the estimates for domain discontinuities.
- `RE_s`: A diagonal covariance matrix with equal variances for the regular and alternative approach, i.e.: $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_2$.

3.3 Univariate Fay-Herriot model for domain discontinuities

The univariate FH model for the direct estimates of the discontinuities starts with defining a measurement error model for the GREG estimates of the discontinuities:

$$\hat{\Delta}_i = \Delta_i + z_i \tag{3.8}$$

with $\Delta_i = y_i^r - y_i^a$ the true discontinuity of domain i under a complete enumeration of the population under both approaches, $\hat{\Delta}_i = \hat{y}_i^r - \hat{y}_i^a$ the GREG estimate for Δ_i based on the parallel run and $z_i = e_i^r - e_i^a$ the sampling error of $\hat{\Delta}_i$. It is assumed that $z_i \simeq \mathcal{N}\left(0, \psi_i^r + \psi_i^a\right)$ and that the design variances of the sampling errors are known. For Δ_i the following linear model is assumed:

$$\Delta_i = \mathbf{x}_i^t \boldsymbol{\beta} + v_i, \tag{3.9}$$

with $v_i \simeq \mathcal{N}\left(0, \sigma_v^2\right)$. Inserting (3.9) into (3.10) gives:

$$\hat{\Delta}_i = \mathbf{x}_i^t \boldsymbol{\beta} + v_i + z_i. \tag{3.10}$$

3.4 Estimation of the bivariate Fay-Herriot model

The models developed in Subsections 3.2 and 3.3 are fitted with a HB approach using Markov Chain Monte Carlo (MCMC) sampling. In particular the Gibbs sampler is used. The following priors are used for the model parameters and hyperparameters. For the regression coefficients uniform improper priors are assumed, i.e., $\boldsymbol{\beta} \sim 1$. For the random domain effects a multivariate normal prior is used: $\mathbf{v} \,|\, \boldsymbol{\Sigma} \simeq \mathcal{N}\left(\mathbf{0}_{2m}, \mathbf{I}_m \otimes \boldsymbol{\Sigma}\right)$.

In the case of a full covariance matrix for the random domain effects, the prior for $\boldsymbol{\Sigma}$ is taken to be a scaled inverse Wishart distribution (O'Malley and Zaslavsky, 2008). This distribution is obtained by writing $\boldsymbol{\Sigma} = \text{Diag}(\boldsymbol{\xi})\,\tilde{\boldsymbol{\Sigma}}\,\text{Diag}(\boldsymbol{\xi})$, with $\boldsymbol{\xi} = \left(\xi^r, \xi^a\right)^t$ and assuming a standard normal distribution for ξ^r and ξ^a, i.e., $\xi^x \simeq \mathcal{N}(0, 1)$, $(x \in (a, r))$ and an inverse Wishart distribution for $\tilde{\boldsymbol{\Sigma}}$, i.e., $\tilde{\boldsymbol{\Sigma}} \simeq \text{Inv} - \text{Wish}\left(v_v, \boldsymbol{\Phi}_v\right)$, with $v_v = d + 1$ degrees of freedom, with d the dimension of $\tilde{\boldsymbol{\Sigma}}$ which is equal to 2 in this application, and scale parameter $\boldsymbol{\Phi}_v = \mathbf{I}_2$. In the case of a diagonal covariance matrix for the random domain effects, the priors for σ_q (in the case of unequal variances) or σ (in the case of equal variances) are half-Cauchy distributions. These are more robust prior distributions than the more commonly used inverse chi-squared distribution (Gelman, 2006). The inverse chi-squared distribution might be informative, even in the case of small scale and shape parameters. In addition convergence problems might occur with the Gibbs sampler. Both problems are largely avoided with a redundant multiplicative parametrization of the random effects (Gelman, 2006; Gelman, Van Dyk, Huang and Boscardin, 2008; Polson and Scott, 2012). See van den Brakel and Boonstra (2018) for more details on the priors of this model.

Let $\hat{\mathbf{y}}$ denote the $2m$ column vector obtained by stacking the m column vectors $\hat{\mathbf{y}}_i$, and $\mathbf{X}$ the matrix obtained by stacking the matrices $\mathbf{X}_i$. In the case of unequal regression coefficients, $\mathbf{X}$ is a $2m \times 2p$ matrix. In the case of equal regression coefficients, $\mathbf{X}$ is a $2m \times p$ matrix. The likelihood function can be written as

$$p\left(\hat{\mathbf{y}} \,|\, \boldsymbol{\theta}\right) = \mathcal{N}\left(\mathbf{X}\boldsymbol{\beta} + \mathbf{v}, \boldsymbol{\Psi}\right), \tag{3.11}$$

with $\boldsymbol{\Psi} = \oplus_{i=1}^{m} \boldsymbol{\Psi}_i$ a $2m \times 2m$ diagonal matrix with the design variances of the direct estimates $\hat{\mathbf{y}}$ and $\boldsymbol{\theta}$ a vector containing all model parameters. The joint prior distribution $p(\boldsymbol{\theta})$ equals the product of the aforementioned priors. The posterior distribution of $\boldsymbol{\theta}$ is proportional to the joint density, i.e., $p\left(\boldsymbol{\theta} \,|\, \hat{\mathbf{y}}\right) \propto p(\boldsymbol{\theta})\, p\left(\hat{\mathbf{y}} \,|\, \boldsymbol{\theta}\right)$. The model is fitted using the Gibbs sampler Geman and Geman (1984);

Gelfand and Smith (1990). The full conditional distributions used in the Gibbs sampler are specified in van den Brakel and Boonstra (2018).

For each model considered, the Gibbs sampler is run in three independent chains with randomly generated starting values. The length of each chain after the burn-in period for each run is 10,000 iterations. This gives 30,000 draws to compute estimates and standard errors. The convergence of the MCMC simulation is assessed using trace and autocorrelation plots as well as the Gelman-Rubin potential scale reduction factor (Gelman and Rubin, 1992), which diagnoses the mixing of the chains. The diagnostics suggest that all chains converge well within 500 draws. The estimated Monte Carlo simulation errors are small compared to the posterior standard errors for all parameters, so that the number of draws are more than sufficient for our purposes.

The estimands of interest are expressed as functions of the parameters, and applying these functions to the MCMC output for the parameters results in draws from the posteriors for these estimands. Domain predictions for the target variables under the bivariate FH model are obtained as the posterior means approximated by the Gibbs sampler output and are denoted as $\tilde{y}_i^{q,b\text{FH}}$. Domain predictions for the discontinuities are obtained as the posterior means of (3.6) or (3.7) approximated by the Gibbs sampler output and are denoted as $\tilde{\Delta}_i^{b\text{FH}}$. Mean squared errors for $\tilde{y}_i^{q,b\text{FH}}$ and $\tilde{\Delta}_i^{b\text{FH}}$ are obtained as posterior variances approximated from the Gibbs sampler output.

The methods are implemented in R using the `mcmcsae` R-package (Boonstra, 2020).

3.5 Pooling design variances

Estimates for the design variances ψ_i^r and ψ_i^a are available from the GREG estimator and are used as if the true design variances are known. This is a standard assumption in small area estimation. Therefore it is important to provide reliable estimates for these design variances. For the regular survey the variance estimates of the GREG estimates are considered to be reliable enough to be used in the FH model. For the alternative survey the estimates of the design variances are unreliable and therefore smoothed to improve their stability of the estimates of ψ_i^a. Under the assumption that the population variances of the GREG residuals under the alternative approach are equal accross domains, the analysis-of-variance type of pooled variance estimator is used:

$$\psi_i^a = \frac{1 - f_i^a}{n_i^a} \frac{1}{n^a - m} \sum_{i=1}^{m} (n_i^a - 1)\, S_{i;\text{GREG}}^{a^2}$$

with f_i^a the sample fraction in domain i of the alternative survey, n_i^a the number of respondents in domain i under the alternative survey, $n^a = \sum_{i=1}^{m} n_i^a$, and $S_{i;\text{GREG}}^{a^2}$ the estimated population variance of the GREG residuals.

Alternatively, variance estimates of the direct estimates can be smoothed by modeling the variance estimates along with the GREG estimates themselves, (You and Chapman, 2006) and Sugasawa, Tamae and Kubokawa (2017). Their approach can be traced back to Arora and Lahiri (1997). Another possibility

is to smooth variance estimates by applying generalized variance functions, (Wolter (2007), Chapter 7, and Hawala and Lahiri (2018)).

3.6 Model selection and evaluation

Frequently applied model selection criteria in HB settings are the Widely Applicable Information Criterion or Watanabe-Akaike Information Criteria (WAIC) (Watanabe, 2010, 2013) and the Deviance Information Criteria (DIC) (Spiegelhalter, Best, Carlin and van der Linde, 2002). They are popular because they are easy to compute from MCMC simulation output and because of their ability to make a reasonable tradeoff between model fit and model complexity. The WAIC is seen as an improvement on the DIC since the latter can produce negative estimates for the effective number of parameters and it is not defined for singular models (Vehtari, Gelman and Gabry, 2017). The penalty used for model complexity in DIC and WAIC is closely related to the effective number of parameters proposed by Hodges and Sargent (2001) for linear multilevel models where each fixed effect contributes one degree of freedom and the random effects contribute a value in the range between zero and m, depending on the size of the variance component. As follows from the definition of WAIC, models with lower WAIC values are preferred. The WAIC estimates are uncertain and an approximation of its standard error is provided by Vehtari et al. (2017) equation (23) and can be computed using R package `loo` (Vehtari, Gelman and Gabry, 2015).

Covariates are selected from the set of auxiliary variables listed in van den Brakel and Boonstra (2018) using a step-forward selection procedure. Various models are compared using the aforementioned WAIC estimates. From the set of potential covariates, the covariate with the lowest WAIC value is selected in the model. This selection process is iteratively repeated as long as adding a new covariate further decreases the WAIC value. In this application, this step-forward selection procedure, further abbreviated as `step-WAIC`, often results in models with a large number of covariates. Since the WAIC values are estimates that contain error, it appears that it might not be desirable to minimize the WAIC by adding covariates to the model as long as it reduces the point estimates of the WAIC. As an alternative we applied a step-forward selection procedure where covariates are added to the model as long as a new covariate decreases the WAIC with a value that exceeds the estimated standard error of the WAIC. This method will be referred to as `step-WAIC-se`.

The step-forward selection procedure is applied to each of the six different combinations of the two fixed effect versions (`FE_uq` and `FE_eq`) and the three covariance structures of the random component (`RE_f`, `RE_d`, and `RE_s`). From the resulting six models the one with the lowest WAIC value is selected. Model adequacy of these six selected models is evaluated with posterior predictive checks. This implies that replicate data sets, simulated from the posterior predictive distribution are compared with the originally observed data to study systematic discrepancies and to evaluate how well the selected model fits the observed data (Gelman, Carlin, Stern, Dunson, Vehtari and Rubin, 2004). Posterior predictive p-values are calculated for six different tests that evaluate particular aspects of the posterior predictive

distribution. Posterior predictive p-values for the domain discontinuities are defined as $p = P\left(T\left(\hat{\boldsymbol{\Delta}}^{\text{sim}}, \boldsymbol{\Delta}\right) \geq T\left(\hat{\boldsymbol{\Delta}}, \boldsymbol{\Delta}\right) \middle| \hat{\mathbf{y}}\right)$, where $\hat{\boldsymbol{\Delta}}^{\text{sim}} = \left(\hat{\Delta}_1^{\text{sim}}, \ldots, \hat{\Delta}_m^{\text{sim}}\right)^t$ are replicates of the observed discontinuities for the m domains under the posterior predictive distribution, $\hat{\boldsymbol{\Delta}} = \left(\hat{\Delta}_1, \ldots, \hat{\Delta}_m\right)^t$ the observed direct estimates for the m domain discontinuities and $T(\hat{\boldsymbol{\Delta}}^{\text{sim}}, \boldsymbol{\Delta})$ (or $T(\hat{\boldsymbol{\Delta}}, \boldsymbol{\Delta})$) a test statistic that depends on $\hat{\boldsymbol{\Delta}}^{\text{sim}}$ (or $\hat{\boldsymbol{\Delta}}$) and unknown true values for the m domain discontinuities $\boldsymbol{\Delta} = (\Delta_1, \ldots, \Delta_m)^t$. Posterior predictive p-values are estimated from the Gibbs sampler output as the average over the S Monte Carlo samples

$$\hat{p} = \frac{1}{S} \sum_{s=1}^{S} I\left(T\left(\hat{\boldsymbol{\Delta}}^s, \boldsymbol{\Delta}^s\right) \geq T\left(\hat{\boldsymbol{\Delta}}, \boldsymbol{\Delta}^s\right)\right), \tag{3.12}$$

with $I(A)$ the indicator function with value one if the condition A is fulfilled and zero otherwise, $\hat{\boldsymbol{\Delta}}^s = \left(\hat{\Delta}_1^s, \ldots, \hat{\Delta}_m^s\right)^t$ the m observed domain discontinuities in the s^{th} replicate of the MCMC simulation and $\boldsymbol{\Delta}^s = \left(\Delta_1^s, \ldots, \Delta_m^s\right)^t$ the true values of the m domain discontinuities in the s^{th} replicate of the MCMC simulation. If a model fits the observed data adequately, then it is expected that $T\left(\hat{\boldsymbol{\Delta}}, \boldsymbol{\Delta}^s\right)$ is in the bulk of the histogram of the replicates $T\left(\hat{\boldsymbol{\Delta}}^s, \boldsymbol{\Delta}^s\right)$. Therefore p-values close to zero or one are indications of a poor fit with respect to that test statistic. In the expressions below, it is understood that T_x, for $x = 1, \ldots, 6$, is a function of $\left(\hat{\boldsymbol{\Delta}}^s, \boldsymbol{\Delta}^s\right)$ or $\left(\hat{\boldsymbol{\Delta}}, \boldsymbol{\Delta}^s\right)$, depending on the component that is evaluated in (3.12). The following posterior predictive tests are defined (You, 2008):

1. A general goodness-of-fit test statistic $T_1 = \sum_{i=1}^{m} \left(\hat{\Delta}_i - \Delta_i\right)^2 \Big/ \text{Var}\left(\hat{\Delta}_i \middle| \Delta_i\right)$. Here $\text{Var}\left(\hat{\Delta}_i \middle| \Delta_i\right) = \psi_i^r + \psi_i^a$.
2. $T_2 = \max\left(\hat{\Delta}_i\right)$ and $T_3 = \min\left(\hat{\Delta}_i\right)$ which are sensitive for deviations in the tails of the distribution.
3. $T_4 = \frac{1}{m} \sum_{i=1}^{m} \hat{\Delta}_i \equiv \bar{\hat{\Delta}}$, i.e., the mean which is sensitive for bias in the domain predictions.
4. $T_5 = \frac{1}{m-1} \sum_{i=1}^{m} \left(\hat{\Delta}_i - \bar{\hat{\Delta}}\right)^2$, i.e., the variance of the domain estimates, which is sensitive for e.g., overshrinkage.
5. $T_6 = \left|\max\left(\hat{\Delta}_i\right) - \bar{\Delta}\right| - \left|\min\left(\hat{\Delta}_i\right) - \bar{\Delta}\right|$, with $\bar{\Delta} = \frac{1}{m} \sum_{i=1}^{m} \Delta_i$ which is sensitive to asymmetry in the distribution.

4. Results

4.1 Model selection

In Subsection 3.2, two different versions for the fixed effects (`FE_uq` and `FE_eq`) and three different covariance structures of the random effects (`RE_f`, `RE_d` and `RE_s`) are considered for the bivariate FH model. The step-forward selection procedure from Subsection 3.6 is applied to each of these six combinations separately to select covariates. Recall from Subsection 3.1 that for the bivariate FH model

and the univariate FH model for the discontinuities, potential covariates are available from the Municipal Base Administration and the Police Register of Reported Offences. Names of these covariates start with `MBA_` and `PR_` respectively. For the univariate FH model for the alternative survey, direct estimates from the regular CVS are also considered as covariates (van den Brakel et al., 2016). Names of these covariates start with `CVSR_`. See the appendix for an overview of the covariates.

The finally selected models for the bivariate FH model are summarized in Table 4.1. The models presented in Table 4.1 are selected with the `step-WAIC-se` procedure. The `step-WAIC` procedure selects models with a substantially larger amount of covariates which improve the WAIC only marginally. For `offtot` and `unsafe` the `step-WAIC` result in a model with 4 covariates with unequal regression coefficients and diagonal covariance matrices for the random effects (`RE_d` and `RE_s`). For `satispol` and `propvict` the `step-WAIC` result in a model with respectively 3 and 2 covariates with unequal regression coefficients, also with diagonal covariance matrices with equal variances for the random effects (`RE_s`). With only 25 domains, there is a substantial risk that these models overfit the data. An exception is `nuisance`, where both selection procedures result in the same model.

With the `step-WAIC-se` procedure more parsimonious models are obtained as follows from Table 4.1. For total offences, `offtot` and `nuisance`, a model with only one covariate with equal regression coefficients for both surveys (`FE_eq`) is obtained in combination with a full covariance matrix (`RE_f`) with large random domain effects with a strong positive correlation of 0.98 for `offtot` and 0.81 for `nuisance`. Also for `unsafe` a more parsimonious model, with one covariate and equal regression coefficients (`FE_eq`) is obtained with the `step-WAIC-se` procedure. In this case a diagonal covariance matrix with equal variances (`RE_s`) is selected. For `propvict` and `satispol` the `step-WAIC-se` procedure avoids the selection of large amounts of covariates, found with the `step-WAIC` approach. The selected model for `propvict` has a full covariance matrix with a weak positive correlation of 0.1 (`RE_f`), and one covariate with unequal regression coefficients (`FE_uq`). The model for `satispol` has a diagonal covariance matrix with equal variances (`RE_s`) and one covariate with unequal regression coefficients (`FE_uq`). Since parsimonious models are preferred in this application, the models obtained with the `step-WAIC-se` approach are finally selected. See van den Brakel and Boonstra (2018) for a more detailed discussion of the model selection resulting in the finally selected models.

The models selected with the univeriate FH model for the direct estimates of the discontinuities, developed in Subsection 3.3, are summarized in Table 4.2. The models are selected with the `step-WAIC-se` procedure. For `unsafe` the `step-WAIC` results in a model with four covariates. For the other variables the same models are selected as with the `step-WAIC-se` procedure. The univariate FH models developed in van den Brakel et al. (2016) for the alternative survey approach are summarized in Table 4.3.

Standard model diagnostics test the underlying assumptions that the random domain effects and the residuals are normally and independently distributed. Since the number of domains in this application is

small, the power of the tests for normality are weak and do not indicate deviations from normality. Therefore the posterior predictive tests as summarised in Subsection 3.5 are used to evaluate the model adequacy. In addition, the domain predictions aggregated to the national level are compared with the direct estimates at the national level to evaluate the bias introduced with the small area estimation procedures in Subsection 4.2. The posterior predictive p-values for the domain estimates of the target variables and the discontinuities are summarized in Table 4.4 for the bivariate FH model and Table 4.5 for the univariate FH model for the discontinuities. The general measure for goodness-of-fit (T_1) indicates that the fit for the discontinuities of `offtot` is of reduced quality (other models considered had similar high values). The values for the bivariate FH model are slightly better compared to the univariate FH model for the discontinuities. The posterior predictive p-values for maximum (T_2) and minimum (T_3) values do not indicate problems with the tails of the distributions. For these posterior predictive p-values there are no systematic differences between bivariate and univariate FH model. The values for T_4, T_5, and T_6 for the discontinuities of the bivariate model are comparable with the values for the univariate model. The posterior predictive values for the mean (T_4) and asymmetry of the distribution (T_6) indicate that the distributions are symmetrically concentrated around their mean. The posterior predictive p-values for the variance (T_5) indicate some undershrinkage for the discontinuities of `nuisance`, `propvict`, and `offtot` under both the bivariate and univariate FH model.

Table 4.1
Final models bivariate FH model selected with `step-WAIC-se`. All models contain an intercept. ρ: correlation between the random effects

Variable	Model	Covariance structure random effects			
		type	σ_r^2	σ_a^2	ρ
offtot	FE_eq: δ_i + PR_weapon	RE_f	8.77	5.32	0.98
unsafe	FE_eq: δ_i + PR_propcrim	RE_s	1.17	1.17	-
nuisance	FE_eq: δ_i + MBA_immigrnw	RE_f	0.20	0.14	0.81
satispol	FE_uq: MBA_immigr	RE_s	0.78	0.78	-
propvict	FE_uq: PR_propcrim	RE_f	0.79	0.39	0.2

Table 4.2
Final models univariate FH model for direct estimates of the discontinuities. All models contain an intercept

Variable	Model	Variance random effects (σ_v^2)
offtot	PR_propcrim	0.80
unsafe	MBA_benefit	1.03
nuisance	PR_threat	0.038
satispol	MBA_benefit	0.928
propvict	PR_assault	0.485

Table 4.3
Final models univariate FH model for alternative CVS from van den Brakel et al. (2016). All models contain an intercept

Variable	Model	Variance random effects (σ_v^2)
offtot	CVSR_victim	0.003
unsafe	CVSR_nuisance + MBA_benefit + PR_propcrim + PR_drugs	2.997
nuisance	CVSR_nuisance + MBA_old	0.805
satispol	CVSR_funcpol	4.995
propvict	PR_propcrim + MBA_old	7.725

Table 4.4
Posterior predictive *p*-values for the final multivariate FH models from Table 4.1

	T_1	T_2	T_3	T_4	T_5	T_6
Variable	**Discontinuities**					
offtot	0.980	0.797	0.069	0.337	0.968	0.416
unsafe	0.343	0.841	0.833	0.454	0.437	0.912
nuisance	0.927	0.940	0.034	0.345	0.988	0.465
satispol	0.772	0.595	0.392	0.610	0.762	0.484
propvict	0.925	0.261	0.029	0.258	0.970	0.070
	Target variables					
offtot	0.859	0.249	0.024	0.317	0.524	0.089
unsafe	0.308	0.779	0.492	0.420	0.474	0.708
nuisance	0.766	0.317	0.108	0.433	0.504	0.156
satispol	0.742	0.929	0.584	0.457	0.875	0.797
propvict	0.695	0.339	0.168	0.379	0.655	0.194

Table 4.5
Posterior predictive *p*-values for the final univeriate FH models for direct estimates of the discontinuities from Table 4.2

	T_1	T_2	T_3	T_4	T_5	T_6
Variable	**Discontinuities**					
offtot	0.985	0.828	0.071	0.390	0.972	0.438
unsafe	0.382	0.885	0.816	0.464	0.523	0.920
nuisance	0.970	0.941	0.072	0.434	0.978	0.554
satispol	0.814	0.607	0.378	0.607	0.783	0.488
propvict	0.946	0.272	0.052	0.383	0.963	0.092

4.2 Estimation results

In this Subsection estimation results for the three different modelling approaches are discussed. In Subsection 4.2.1 the HB predictions for the target variables under the regular and alternative survey obtained with the bivariate FH model are compared with the direct estimates and with the domain predictions obtained with the univariate FH model where the direct estimates of the regular approach are

potential auxiliary variables in the model selection. Subsequently results for the domain discontinuities are discussed in Subsection 4.2.2. Here the results obtained with the univariate FH model for the discontinuities are also discussed.

With model-based small area estimation, the design variance of the direct estimators is reduced at the cost of accepting some amount of design bias. To evaluate differences in the direct point estimates and the small domain predictions, the following two measures are defined. The first one is the Mean Relative Difference (MRD), which summarizes the differences between the direct estimates and the domain predictions:

$$\text{MRD} = \frac{100\%}{m} \sum_{i=1}^{m} \frac{\hat{y}_i^q - \tilde{y}_i^q}{\hat{y}_i^q},\ q = r,\ a, \tag{4.1}$$

and $\tilde{y}_i^q$ is the domain prediction based on the bivariate FH model or the univariate FH model. The second measure is the Absolute Mean Relative Difference (AMRD) between the direct estimate and the domain prediction, which is defined as:

$$\text{AMRD} = \frac{100\%}{m} \sum_{i=1}^{m} \left| \frac{\hat{y}_i^q - \tilde{y}_i^q}{\hat{y}_i^q} \right|,\ q = r,\ a, \tag{4.2}$$

the increased precision of the small domain predictions is measured with Mean Relative Difference of the Standard Errors (MRDSE) between the direct estimates and the domain predictions and is defined as

$$\text{MRDSE} = \frac{100\%}{m} \sum_{i=1}^{m} \frac{\text{SE}\left(\hat{y}_i^q\right) - \text{SE}\left(\tilde{y}_i^q\right)}{\text{SE}\left(\hat{y}_i^q\right)},\ q = r,\ a, \tag{4.3}$$

these measures are defined in a similar way for the estimates and predictions of the domain discontinuities $\hat{\Delta}_i$ and $\tilde{\Delta}_i$.

4.2.1 Results for variables under the regular and alternative survey

In Table 4.6 the domain predictions and their standard errors averaged over the domains as well as the MRD, AMRD and MRDSE are given for the alternative survey under the univariate FH model with the models presented in Table 4.3. Results under the bivariate FH model, based on the final models of Table 4.1, are presented in Table 4.7 for the variables under the alternative survey and in Table 4.8 for the variables under the regular survey. Comparing the standard errors (SE) and the MRDSE in Table 4.6 and Table 4.7 shows that the bivariate FH model results in stronger reductions of the standard errors for all variables with the exception of `nuisance`. This comes at the cost of an increased bias. Comparing MRD and AMRD in both tables shows that the deviations between the direct estimates and the small area predictions are larger under the bivariate FH model. Comparing the SE and MRDSE in Tables 4.7 and 4.8 shows that the improvement in precision with the bivariate FH model for the regular survey is smaller, as expected since the sample size of the regular survey is larger. The bias in the bivariate FH model

predictions for the regular survey are also smaller, which follows from a comparison of MRD and AMRD in Tables 4.7 and 4.8.

The domain predictions under the univariate and bivariate FH model are plotted against the GREG estimates in Figures 4.1 through 4.5. The graphs also contain the GREG estimate at the national level versus the domain predictions aggregated to the national level according to (4.4). Figures 4.1 and 4.3 show that there is only a small amount of shrinkage for `offtot` and `nuisance`. Figure 4.2 shows for `unsafe` that the bivariate FH model shrinks the domain predictions for the alternative survey while the amount of shrinkage for the univariate FH model for the alternative CVS and the bivariate FH model for the regular survey is smaller. For `propvict`, see Figure 4.4, there is a small difference between the amount of shrinkage of the alternative CVS under the bivariate and univariate model. From Figure 4.5 it follows that the bivariate FH model for `satispol` cannot adequately model the observations under the alternative survey with the auxiliary information from the two registers (MBA and PRRO). In this case the domain predictions of `satispol` under the alternative approach display extreme overshrinkage. The univariate FH model indeed selects the same auxiliary variable from the regular survey only, see Table 4.3 and results in more realistic domain predictions.

For variables related to opinions and views such as `unsafe` and `satispol`, the reduction in the standard errors is accompanied by a relatively strong increase in the bias. This is especially the case with the small area prediction of the bivariate FH model for the alternative survey. For these variables, there are no strongly correlated covariates in the MBA and PRRO. In these cases the univariate FH model performs better since related covariates from the regular survey are selected (see Table 4.3), while the bivariate model doesn't detect correlation between the random effects (see Table 4.1).

Table 4.6
Average of domain predictions alternative survey with univariate FH model from van den Brakel et al. (2016)

Variable	HB est.	SE	MRD (%)	AMRD (%)	MRDSE (%)
offtot	33.21	2.90	-0.44	7.03	47.74
unsafe	19.83	1.64	-0.96	7.58	41.16
nuisance	1.29	0.08	-0.74	5.02	37.96
satispol	55.09	2.54	-0.11	6.43	61.98
propvict	9.85	0.84	-3.17	11.86	60.69

Table 4.7
Average of domain predictions alternative survey with bivariate FH model

Variable	HB est.	SE	MRD (%)	AMRD (%)	MRDSE (%)
offtot	33.26	2.82	-0.99	6.93	49.36
unsafe	19.82	1.21	-2.54	11.97	56.47
nuisance	1.28	0.08	-0.98	4.28	35.72
satispol	55.08	1.97	-0.49	8.97	70.06
propvict	9.91	0.73	-4.81	14.70	65.35

Table 4.8
Average of domain predictions regular survey with bivariate FH model

Variable	HB est.	SE	MRD (%)	AMRD (%)	MRDSE (%)
offtot	41.34	3.76	0.96	4.56	17.95
unsafe	24.22	1.07	-0.01	6.02	46.78
nuisance	1.60	0.09	0.38	2.62	15.93
satispol	60.82	1.47	-0.77	5.38	64.83
propvict	12.18	0.88	1.58	7.84	43.70

The direct estimates at the national level are accurate estimates since they are based on sufficiently large sample sizes. Therefore the bias in model-based domain predictions is often assessed by comparing the direct estimates at the national level with the domain predictions aggregated to the national level. The target variables in this application are all defined as population means. Therefore the aggregated domain predictions are obtained as the average over the domains weighted with the relative domain sizes,

$$\tilde{y}^q = \sum_{i=1}^{m} \frac{N_i}{N} \tilde{y}_i^q \tag{4.4}$$

with N_i the population size of domain i and N the size of the total population.

Table 4.9 compares the weighted average of the domain predictions according to (4.4) with the national GREG estimates. For the univariate FH model for the alternative CVS, the aggregated domain predictions are almost exactly equal to the GREG estimates at the national level. For the bivariate FH model the differences are slightly larger but the aggregated domain predictions are still very close to the GREG estimates at the national level. The largest relative difference amounts to 3% and is observed for offtot under the regular survey.

Table 4.9
GREG estimates national level and aggregated HB predictions regular and alternative survey approach (4.4)

Variable	Regular		Alternative			Discontinuity			
	GREG	biv. FH	GREG	biv. FH	uni. FH	GREG	biv. FH	uni. FH	Δ FH *)
offtot	43.79	42.47	34.09	34.02	34.09	9.7	8.45	9.7	9.04
unsafe	25.07	24.89	20.48	20.49	20.48	4.59	4.40	4.59	4.69
nuisance	1.67	1.66	1.34	1.34	1.34	0.33	0.32	0.34	0.33
satispol	59.88	60.36	55.10	55.06	55.12	4.78	5.29	5.04	5.07
propvict	13.02	12.76	10.32	10.33	10.32	2.70	2.43	2.70	2.63

*) : Δ FH are the HB predictions with univariate FH model for the direct estimates of the discontinuities, weighted similarly to (4.4).

Figure 4.1 Domain estimates GREG versus HB predictions offtot. Upper panel: regular survey using bivariate FH model, middle panel: alternative survey using bivariate FH model, lower panel alternative survey using univariate FH model. Domain predictions are aggregated at the national level according to (4.4).

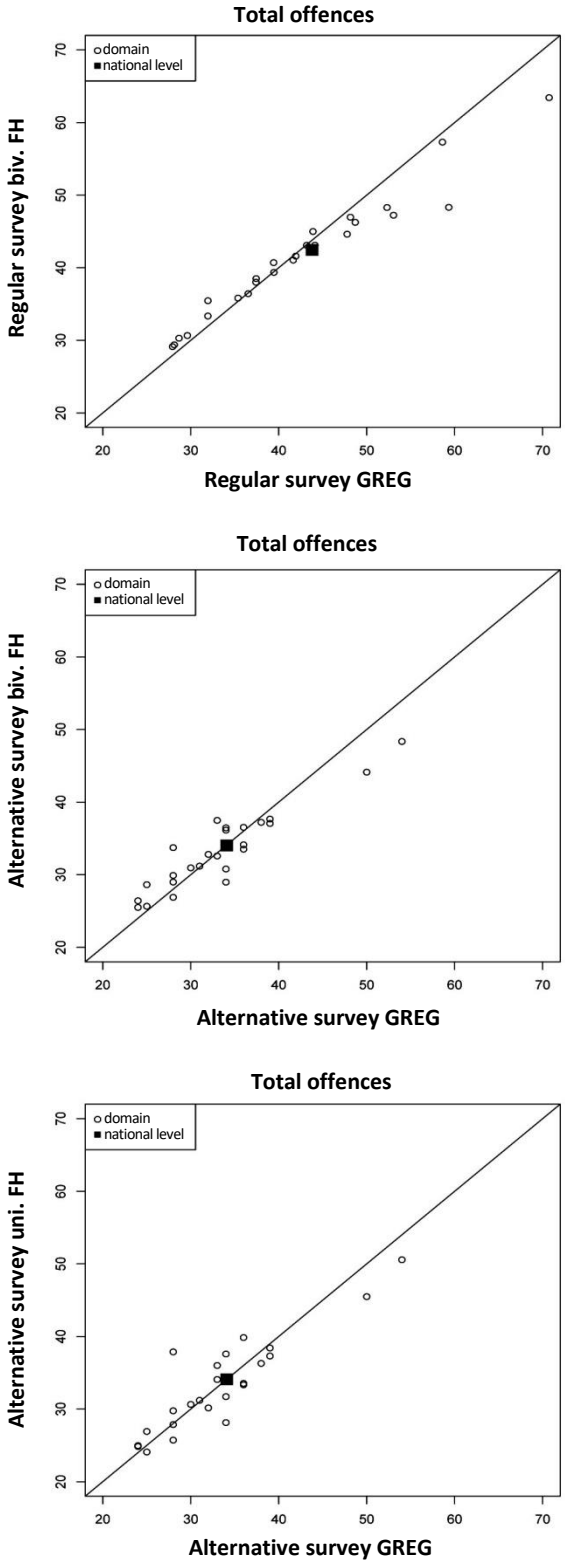


Figure 4.2 Domain estimates GREG versus HB predictions unsafe. Upper panel: regular survey using bivariate FH model, middle panel: alternative survey using bivariate FH model, lower panel alternative survey using univariate FH model. Domain predictions are aggregated at the national level according to (4.4).

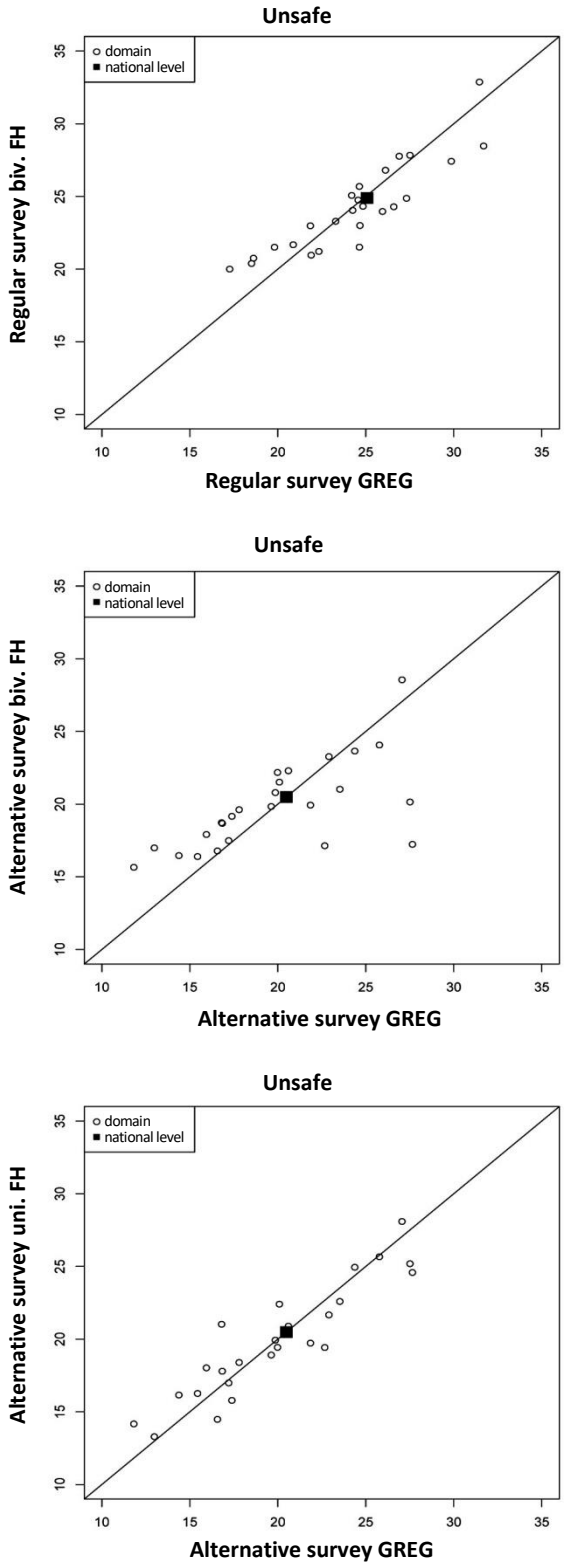


Figure 4.3 Domain estimates GREG versus HB predictions nuisance. Upper panel: regular survey using bivariate FH model, middle panel: alternative survey using bivariate FH model, lower panel alternative survey using univariate FH model. Domain predictions are aggregated at the national level according to (4.4).

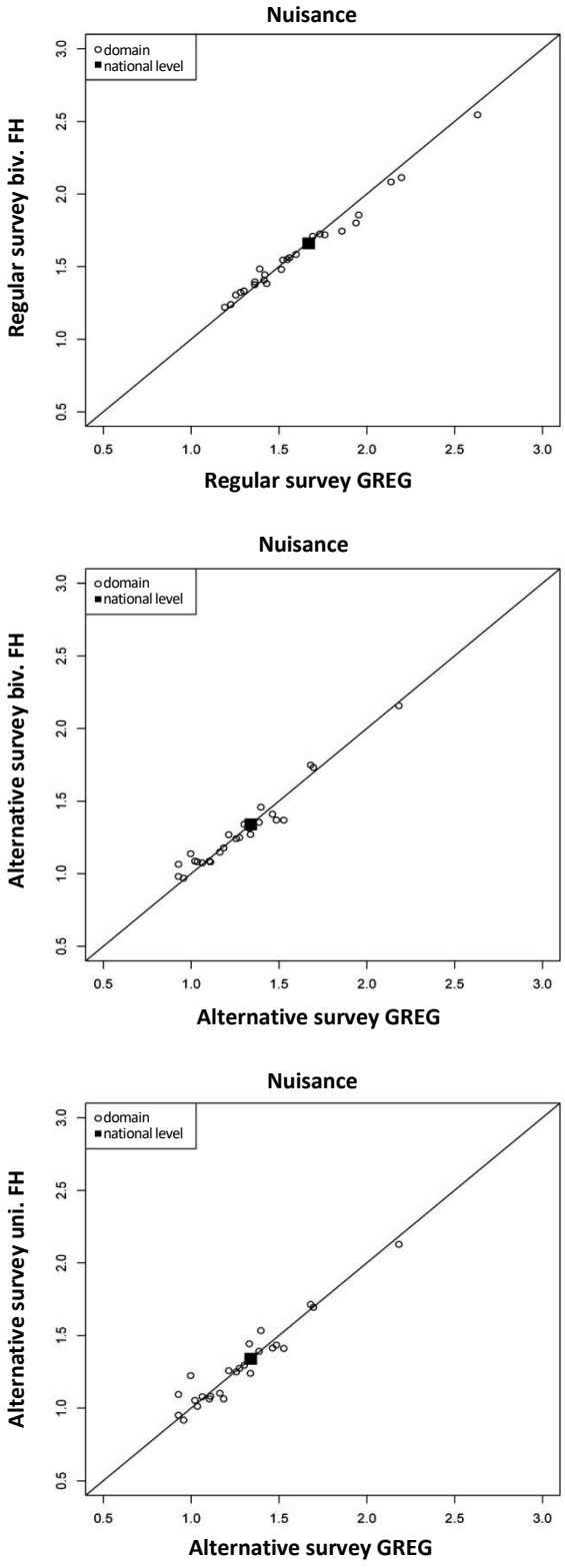


Figure 4.4 Domain estimates GREG versus HB predictions propvict. Upper panel: regular survey using bivariate FH model, middle panel: alternative survey using bivariate FH model, lower panel alternative survey using univariate FH model. Domain predictions are aggregated at the national level according to (4.4).

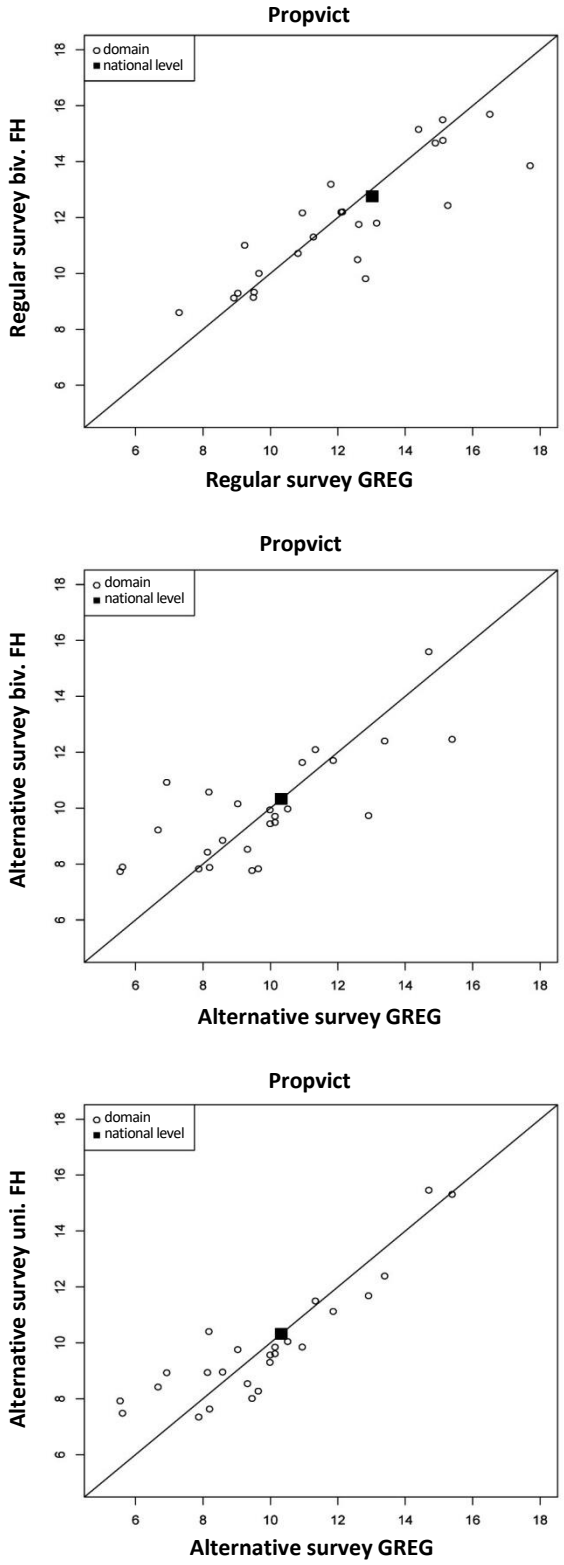
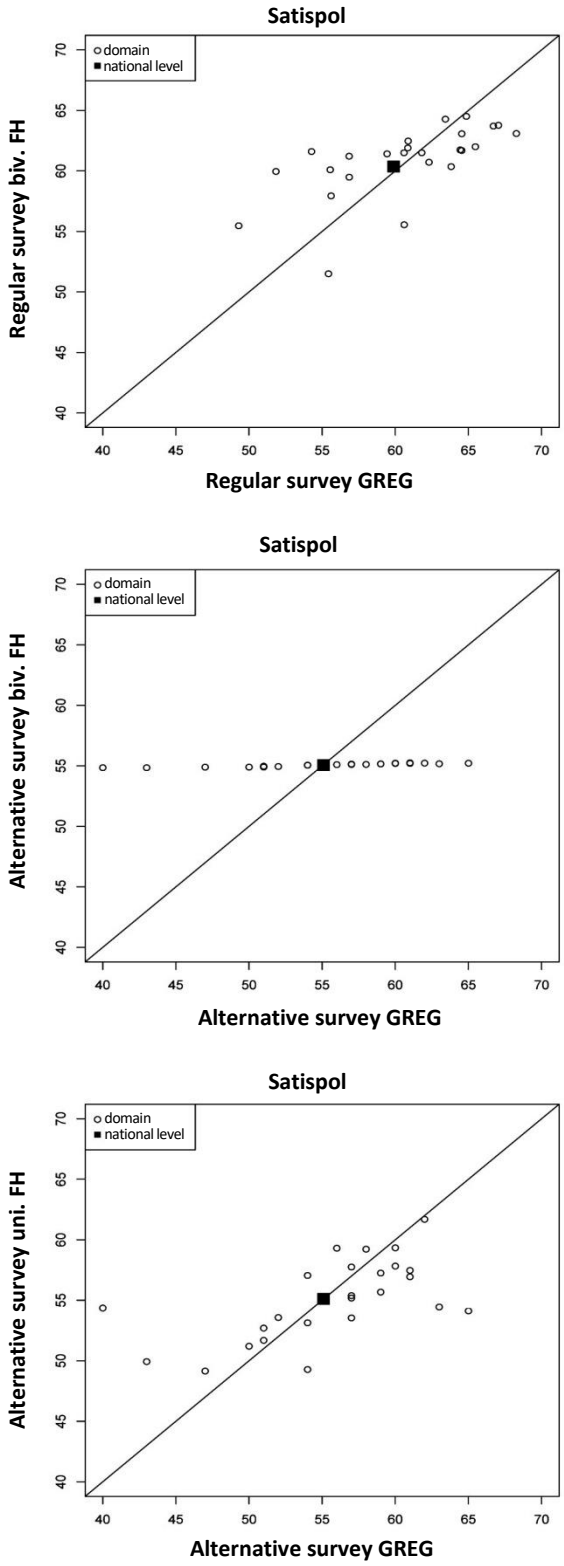


Figure 4.5 Domain estimates GREG versus HB predictions satispol. Upper panel: regular survey using bivariate FH model, middle panel: alternative survey using bivariate FH model, lower panel alternative survey using univariate FH model. Domain predictions are aggregated at the national level according to (4.4).



4.2.2 Results for discontinuity estimates

In the last four columns of Table 4.9, the GREG estimates for the discontinuities at the national level are compared with the domain predictions obtained with the univariate FH model for the alternative CVS, the bivariate FH model and the univariate FH model for the discontinuities, aggregated to the national level using (4.4). The differences between the GREG estimates for the discontinuities at the national level and the aggregated domain predictions are the largest for the bivariate model and the smallest for the univariate FH model for the alternative CVS. This can be expected since the bivariate FH model shrinks both the domain estimates for the regular and alternative survey. With the univariate FH model for the alternative CVS, only the estimates for the alternative survey are replaced by domain predictions, while the estimates for the regular survey are not adjusted. In addition the domain predictions for the alternative survey have larger MRD's and AMRD's under the bivariate FH model compared to the univariate FH model (compare Table 4.6 and 4.7). The differences for the univariate FH model for the discontinuities are smaller compared to the bivariate FH model but larger compared to the univariate FH model for the alternative CVS.

In Tables 4.10, 4.11, and 4.12 the domain predictions and their standard errors for the discontinuities averaged over the domains as well as the MRD and MRDSE are summarized for the univariate FH model for the alternative CVS, bivariate FH model and the univariate FH model for the discontinuities respectively. The MRD's are large because the GREG estimates for the discontinuities in the denominator of (4.11) frequently take values close to zero, which make these indicators unstable. Therefor the AMRD is replaced by the median of the absolute relative differences, $\left|\left(\hat{y}_i^q - \tilde{y}_i^q\right)/\hat{y}_i^q\right|$, and is abbreviated as MARD. The latter are indeed more stable indicators for bias. The MARD is the smallest for the univariate FH model for the alternative CVS, since this approach only adjusts the domain predictions of the alternative CVS. The MARD values for the bivariate FH model on their turn are smaller than those for the univariate FH model for the discontinuities.

With the exception of `nuisance` the standard errors for the domain predictions under the bivariate FH model are smaller compared to the univariate FH model for the alternative CVS. In the case of `propvict` and `offtot`, this is the result of slightly more precise domain predictions for the alternative survey with respect to the univariate FH model (compare Table 4.6 with 4.7), a clear improvement in precision of the domain predictions of the regular survey compared to the GREG estimators (Table 4.8 and 2.2) and the positive correlation between the random effects. In the case of `satispol` and `unsafe` this is mainly the result of a clear improvement of precision of the domain predictions with the bivariate FH model for the regular compared to the GREG estimators (Table 4.8 and 2.2) and also a clear improvement of the precision of the domain predictions with the bivariate FH model for the alternative survey compared to the univariate model (compare Table 4.6 with Table 4.7).

For all five variables, the smallest standard errors are obtained with the univariate FH model for the discontinuities. This comes at the cost of a larger bias, as illustrated with the MARD values. An exception is `satispol`, for which the bias in terms of MARD for the bivariate FH model is clearly larger than the univariate FH model for the discontinuities. For this variable the bias is the lowest with the univariate FH model for the alternative CVS, but the reduction of the standard errors is also smaller.

The last columns of Tables 4.11 and 4.12 contain the shrinkage factors for the domain discontinuities averaged over the domains. For the univariate FH model for the discontinuities the shrinkage factors for the predictions of the domain discontinuities, i.e., the weights attached to the direct estimator for the discontinuities, are defined as $\gamma_i = \hat{\sigma}_v^2 / \left(\hat{\sigma}_v^2 + \left(\psi_i^r + \psi_i^a\right)\right)$. For the bivariate model the shrinkage factors for the predictions of the domain discontinuities are defined as $\gamma_i = \boldsymbol{\iota}\hat{\boldsymbol{\Sigma}}\boldsymbol{\iota}^t / \left(\boldsymbol{\iota}\hat{\boldsymbol{\Sigma}}\boldsymbol{\iota}^t + \left(\psi_i^r + \psi_i^a\right)\right)$, with $\boldsymbol{\iota} = (1, -1)^t$. The average shrinkage factor is defined as $\bar{\gamma} = 1 / M \sum_{i=1}^{M} \gamma_i$. Note that this statistic is not available for the discontinuities obtained with the univariate FH model for the alternative CVS, since under this approach domain discontinuities are obtained as the contrast between the GREG estimate for the regular survey and the domain prediction for the alternative approach. With the exception of `satispol`, the shrinkage factors under the univariate FH model for discontinuities are a factor 10 smaller compared to those of the bivariate FH model. The question rises whether the extremely small shrinkage factors of the univariate FH model for the discontinuities overshrink the direct estimates of the discontinuities.

Table 4.10
Domain predictions for discontinuities univariate FH model for the alternative CVS

Variable	HB est.	SE	MRD (%)	MARD (%)	MRDSE (%)
offtot	9.08	3.92	-5.67	22.14	48.47
unsafe	4.55	2.46	42.98	23.44	29.45
nuisance	0.33	0.07	-5.80	11.49	57.49
satispol	5.52	4.72	99.26	47.72	40.86
propvict	2.70	1.83	-142.60	30.49	32.75

Table 4.11
Domain predictions for discontinuities bivariate FH model

Variable	HB est.	SE	MRD (%)	MARD (%)	MRDSE (%)	$\bar{\gamma}$
offtot	8.07	2.68	1.94	23.34	63.60	0.208
unsafe	4.40	1.56	55.49	41.24	55.09	0.317
nuisance	0.31	0.09	-4.01	18.43	44.47	0.327
satispol	5.74	2.46	228.50	73.43	68.75	0.019
propvict	2.27	1.10	-113.00	27.11	59.02	0.376

Table 4.12
Domain predictions for discontinuities univariate FH model for the direct estimates of the discontinuities

Variable	HB est.	SE	MRD (%)	MARD (%)	MRDSE (%)	$\bar{\gamma}$
offtot	8.35	2.25	-33.39	26.63	69.35	0.012
unsafe	4.43	1.48	45.45	42.14	59.99	0.082
nuisance	0.32	0.06	-13.85	20.97	63.04	0.049
satispol	5.67	2.39	186.07	65.33	69.61	0.014
propvict	2.54	0.91	-55.59	45.73	65.84	0.032

Plots of discontinuities estimated with the GREG estimator, the univariate FH model for the alternative CVS, the bivariate FH model and the univariate FH model for the discontinuities are provided in Figures 4.6 through 4.10. The predictions for the domain discontinuities obtained with the three models are more stable compared to the GREG estimates. This is e.g., clearly illustrated with unsafe (Figure 4.7), where the GREG estimates for the discontinuity are sometimes positive and sometimes negative. The predictions for the domain discontinuities under the bivariate FH model and the univariate FH model for the discontinuities are consistently positive, which appears more plausible since it is unlikely that the domain discontinuities have opposite signs. The predictions for the domain discontinuities under the univariate FH model for the alternative CVS are closer to the GREG estimates and consequently less stable. A similar pattern can be observed for the other variables.

These plots illustrate that for propvict, offtot and unsafe the bivariate FH model results in a clear improvement of the predictions for the domain discontinuities compared to the univariate FH model for the alternative CVS. For nuisance the standard errors for the discontinuities increase with the bivariate FH model compared to the univariate FH model for the alternative CVS. The bivariate FH model for satispol cannot adequately model the observations under the alternative survey with the auxiliary information from the two registers (MBA and PRRO). In this case the domain predictions of satispol under the alternative approach display overshrinkage. The univariate FH model indeed selects an auxiliary variable from the regular survey, see Table 4.3, and clearly performs better.

It was anticipated that it would be difficult to produce reasonable predictions for the domain discontinuities with the univariate FH model for the direct estimates of the discontinuities since it is hard to imagine that the available auxiliary variables from registers like the MBA and PRRO contain good predictors for systematic differences in survey errors. Nevertheless, reasonable results are obtained with this more pragmatic approach. A possible interpretation is that the discontinuities are to some extent proportional to the values of the target variable and therefore show some systematic pattern that can be explained partially with the selected covariates. A point of concern are the very small shrinkage factors under this model, which might be an indication that the model gives too much weight to the synthetic estimator.

Figure 4.6 Discontinuities offtot based on the GREG estimator (upper panel), univariate FH model (second panel), bivariate FH model (third panel) and univariate FH model for direct estimates discontinuities (lower panel) with a 95% confidence interval.

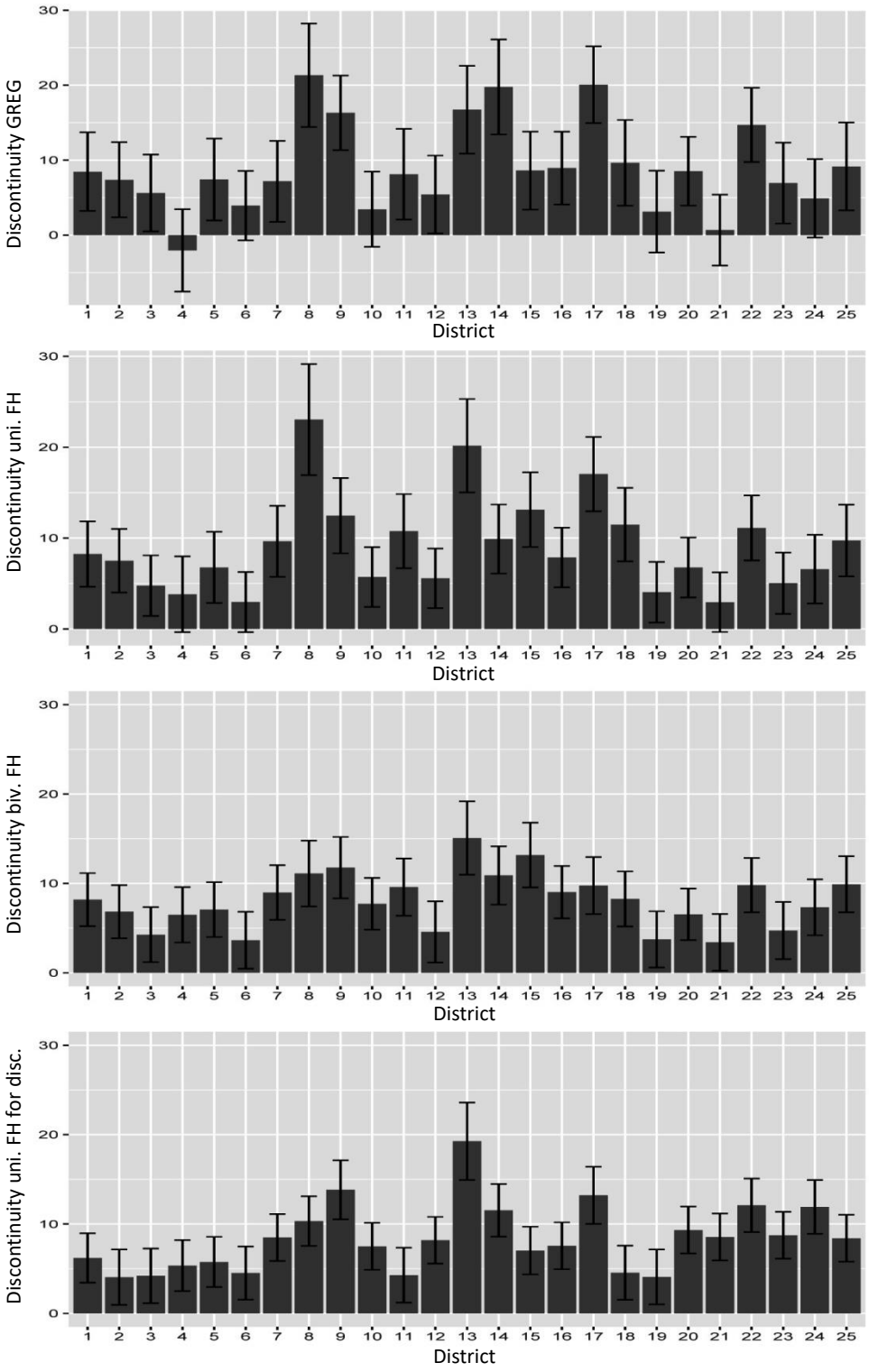


Figure 4.7 Discontinuities unsafe based on the GREG estimator (upper panel), univariate FH model (second panel), bivariate FH model (third panel) and univariate FH model for direct estimates discontinuities (lower panel) with a 95% confidence interval.

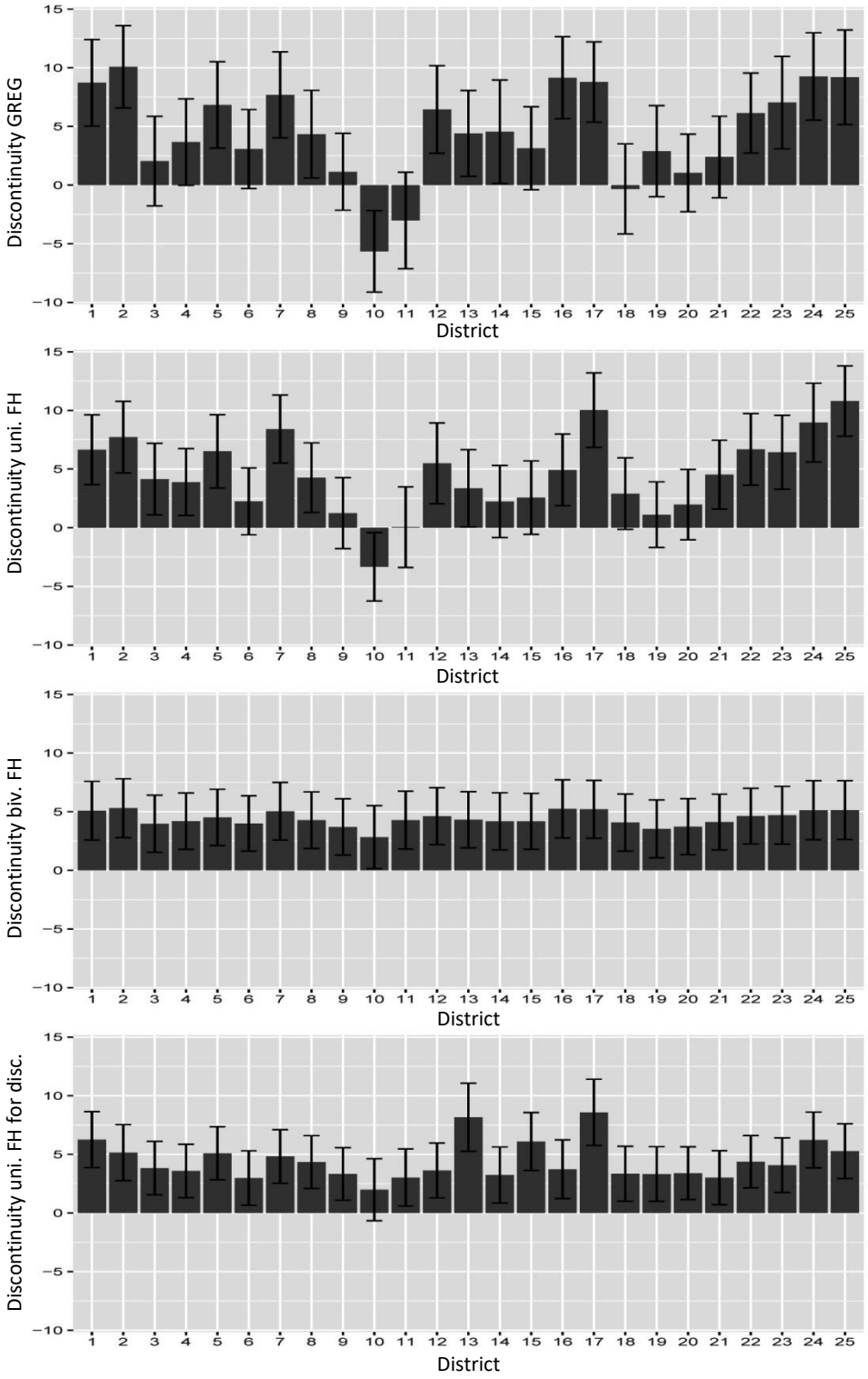


Figure 4.8 Discontinuities nuisance based on the GREG estimator (upper panel), univariate FH model (second panel), bivariate FH model (third panel) and univariate FH model for direct estimates discontinuities (lower panel) with a 95% confidence interval.

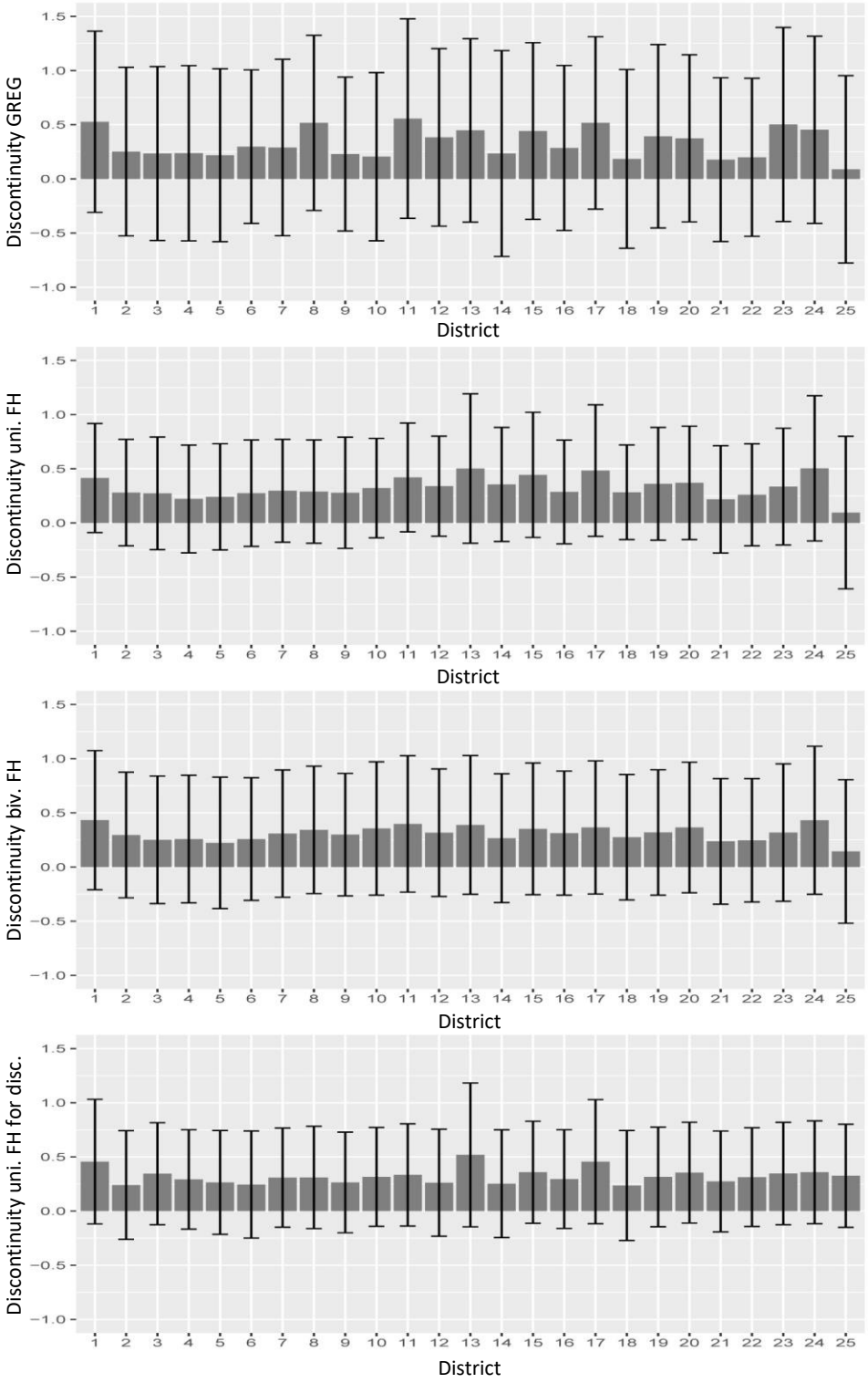


Figure 4.9 Discontinuities propvict based on the GREG estimator (upper panel), univariate FH model (second panel), bivariate FH model (third panel) and univariate FH model for direct estimates discontinuities (lower panel) with a 95% confidence interval.

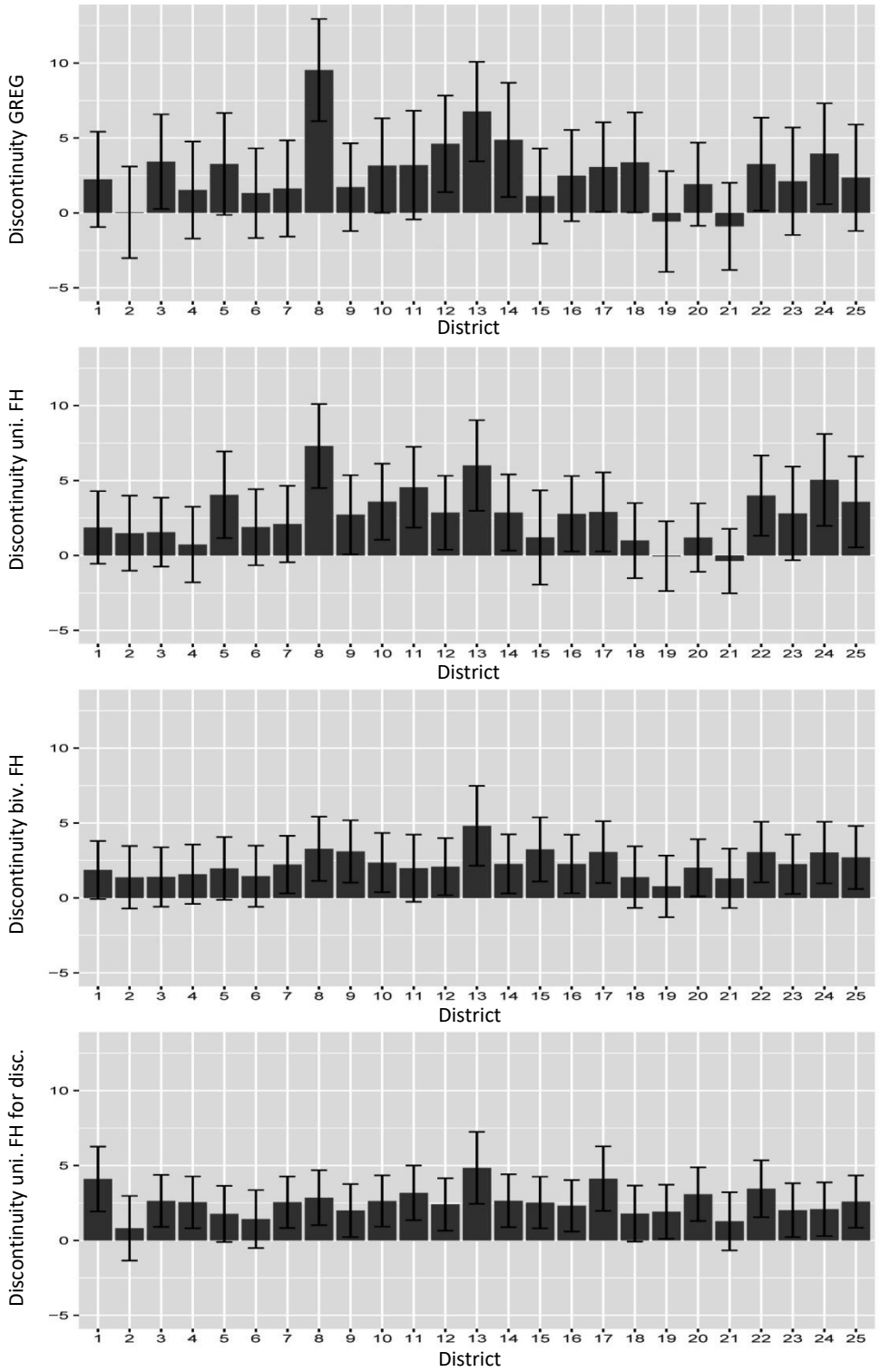
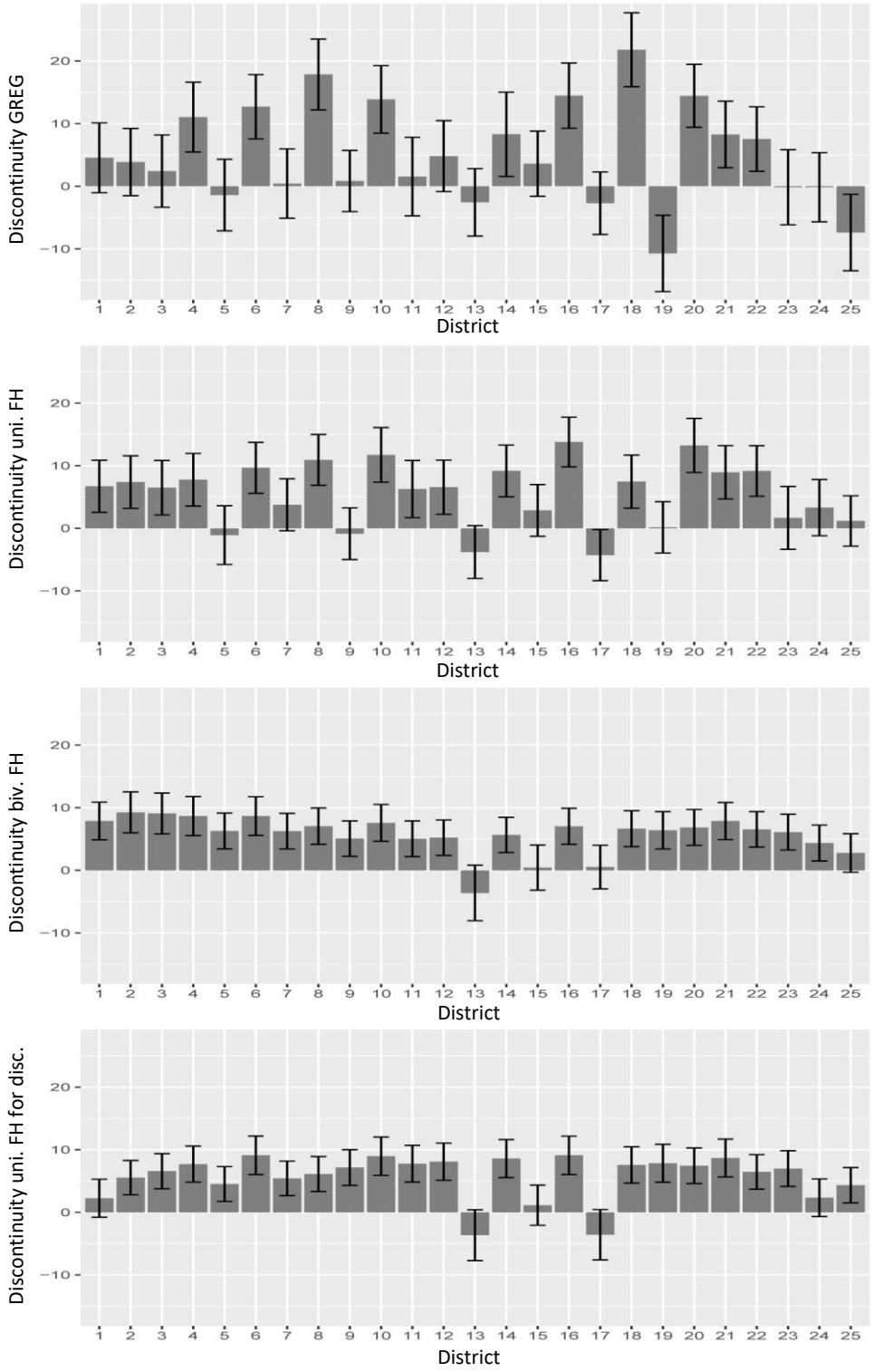


Figure 4.10 Discontinuities satispol based on the GREG estimator (upper panel), univariate FH model (second panel), bivariate FH model (third panel) and univariate FH model for direct estimates discontinuities (lower panel) with a 95% confidence interval.



5. Discussion

Survey process redesigns often result in discontinuities that disturb the comparability of the outcomes over time obtained with a repeated survey. To avoid confounding real period-to-period change with differences in measurement bias, it is important that such discontinuities are quantified during the implementation of a new survey process. A straightforward approach is to collect data under the old and new design in parallel to each other for some period of time. Available budgets for parallel data collection often do not meet the minimum required sample sizes that come from power calculations to detect minimum prespecified differences at certain significance and power levels. This might be sufficient for quantifying discontinuities at the national level but not at the domain level, even for the planned domains of the regular survey. To obtain more precise predictions for the domain discontinuities a small area estimation approaches based on hierarchical Bayesian Fay-Herriot (FH) models is proposed.

In an earlier paper (van den Brakel et al., 2016) a univariate FH model is proposed, where reliable direct domain estimates of the regular survey are considered as potential auxiliary variables in a step-forward model selection procedure to build adequate models for small domain prediction of the small sample assigned to the alternative survey. In this paper a bivariate FH model for the direct estimates obtained under both the regular and alternative survey is proposed as an alternative to obtain adequate predictions for domain discontinuities. In addition a univariate FH model applied to the direct estimates of the discontinuities is considered as a simple alternative. The methods are applied to a small scale parallel run conducted to quantify discontinuities in a survey process redesign of the Dutch Crime Victimization Survey (CVS).

Using direct estimates from the regular survey as auxiliary variables in models for small domains under the alternative approach results in a substantial improvement of precision, compared to univariate models that only use auxiliary variables from available registers. This can be expected since both surveys attempt to measure the same variables with a different survey approach. A drawback of the univariate approach is that the variance estimation procedure for the discontinuities is complex, since a non-negligible covariance between the direct estimates from the regular design and the model based predictions for the alternative design arises. The method is complex since a model-based MSE is combined with a design-based variance of a direct estimator. This might even result in negative variance estimates for the discontinuities. These complications are partially circumvented by developing a design-based estimator for the MSE of the small domain predictions and the covariance component (van den Brakel et al., 2016).

Under a bivariate FH model in a fully Bayesian framework negative variance estimates are avoided since the variances for disontinuities are derived from positive-definite covariance matrices of the bivariate model. The bivariate FH model improves the predictions for the domain discontinuities since the model improves the precision of the estimates of both the regular and alternative approach, and the strong positive correlation between the random domain effects further reduces the variance of the contrasts. For four out of five variables of the Dutch CVS the bivariate FH model indeed resulted in more precise

predictions for domain discontinuities compared to the univariate FH model. Another advantage of the bivariate model is that it improves the domain predictions of both the regular and alternative model while the univariate model assumes that the sample size of the regular survey is sufficiently large to make reliable precise direct domain estimates. The bivariate model is therefore also appropriate in parallel runs where e.g., the sample size of the regular survey is reduced in order to increase the sample size for the alternative survey. Finally the bivariate FH model avoids the complications to account for sampling error in the covariates, which is often required if the direct estimates of the regular survey are used as covariates in a univariate FH model.

For one variable (satisfaction with police performance) no adequate model could be constructed with the available auxiliary variables from the registers only. For this variable the multivariate model seems to result in overshrinkage of the predictions for the domain discontinuities. The results of the univariate model are clearly better in this case since the direct estimates from the regular survey are the only auxiliary variables that result in an adequate model for small domain predictions.

The univariate FH model for the direct estimates of the domain discontinuities turns out to be a reasonable alternative. It avoids the complications of the univariate FH model for the alternative CVS and the method is considerably simpler compared to the bivariate FH model. A point of concern are the extremely small shrinkage factors, which are an indication that the model puts too much weight on the synthetic part of the domain predictions. The bias of these domain predictions is indeed larger compared to that of the bivariate FH model.

A general problem in this application with the step-forward model selection procedure where covariates are included in the model as long as the WAIC value is reduced, is that this results in models with relatively large sets of covariates. With the limited number of domains in this application there is a real risk of overfitting the data. For some variables the covariates appear to be strong predictors for the domain variables, resulting in small random effects. Fitting a model without these covariates results in models with large random effects and strong positive correlations between the regular and alternative survey estimates. For other variables a model with a full covariance structure automatically results in parsimonious models for the fixed effect part, probably because the set of available covariates are less strong predictors for these target variables.

The aformentioned issue of selecting models with too many covariates is circumvented with an alternative step-forward selection approach. Since the WAIC values are estimated from the Gibbs sampler output, these values are observed with some degree of uncertainty. This is an argument not to include covariates if they only result in a small reduction of the WAIC. In an alternative step-forward selection approach, covariates are only selected if the decrease in the WAIC value exceeds the estimated standard error of the WAIC. With this approach parsimoneous models are selected since it avoids the selection of one or more covariates that only marginally improve the WAIC. For variables where initially large sets of covariates were selected, this approach results in a reasonable compromise between model fit and model complexity. As an alternative, models with equal regression coefficients can be considered. Such models

are, however, less appropriate for predicting domain discontinuities if the random effects are small. In such situations the dummy indicator is the only model component that discriminates between the regular and alternative approach. This results in synthetic predictions for domain discontinuities that are almost equal over the domains, and approximately equal to the direct estimator for the discontinuity at the national level. Depending on the type of changes in the survey process, it might be correct to assume that domain discontinuities are equal. In that case the best estimate is obtained with the direct estimator at the national level.

For a better understanding of the properties and behaviour of the three different models for estimating domain discontinuities, including the proposed model selection approach, a comprehensive simulation is required. This will provide a better understanding under what conditions, which of the three different modeling approaches are preferred. Such a study is left for future research.

Acknowledgements

The views expressed in this paper are those of the authors and do not necessarily agree with the policy of Statistics Netherlands. We would like to thank the Associate Editor and the two referees for careful reading of our manuscript and providing useful comments. They proved to be very helpful to improve the quality of our paper.

Appendix

Table A.1
Overview auxiliary data

Variable	Description	Source
MBA_benefit	Percentage of social benefit claimants	MBA
MBA_immigr	Percentage of immigrants in population	MBA
MBA_immigrnw	Percentage of non-western immigrants in population	MBA
MBA_old	Percentage of elderly people (aged over 65)	MBA
MBA_benefit	Percentage of social benefit claimants	MBA
PR_assault	Peported physical assaults	PRRO
PR_propcrim	Property crimes	PRRO
PR_threat	Reported threats	PRRO
PR_weapon	Weapon offences	PRRO
PR_drugs	Illicit drug offences	PRRO
CVSR_nuisance	Perceived nuisance in the neighbourhood	regular survey
CVSR_victim	Percentage of people saying that they have been victim to a crime	regular survey
CVSR_funcpol	Opinion on functioning of the police on a 10-point scale	regular survey

References

Arora, V., and Lahiri, P. (1997). On the superiority of the bayesian method over the blup in small area estimation problems. *Statistica Sinica*, 7, 1053-1063.

Bell, W.R. (1999). Accounting for uncertainty about variances in small area estimation. *Bulletin of the International Statistical Institute*, 52nd Session.

Benavent, R., and Morales, D. (2016). Multivariate Fay-Herriot models for small area estimation. *Computational Statistics and Data Analysis*, 94, 372-390.

Bollineni-Balabay, O., van den Brakel, J.A., Palm, F. and Boonstra, H.J. (2017). Multilevel hierarchical Bayesian vs. state space approach in time series small area estimation: The Dutch Travel Survey. *Journal of the Royal Statisitcal Society*, forthcoming.

Boonstra, H.-J. (2020). `mcmcsae`: *Markov Chain Monte Carlo Small Area Estimation*. R package version 0.5.0.

Boonstra, H.J., and van den Brakel, J.A. (2019). Estimation of level and change for unemployment using structural time series models. *Survey Methodology*, 45, 3, 395-425. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2019003/article/00005-eng.pdf.

Datta, G., Ghosh, M., Nangia, N. and Natarjan, K. (1996). Estimation of median income in four-peron families: A Bayesian approach. In *Bayesian analysis in Statistics and Econometrics*, (Eds., D.A. Berry, K.M. Chaloner and J.M. Geweke), New York: John Wiley & Sons, Inc., 129-140.

Datta, G.S., Lahiri, P., Maiti, T. and Lu, K.L. (1999). Hierarchical Bayes estimation of unemployment rates for the states of the U.S. *Journal of the American Statistical Association*, 94(448), 1074-1082.

Estaban, M.D., Morales, D., Perez, A. and Santamaria, L. (2012). Small area estimation of poverty proportions under area-level time models. *Computational Statistics and Data Analysis*, 56, 2840-2855.

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

Gelfand, A.E., and Smith, A.F.M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398-409.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3), 515-533.

Gelman, A., and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4), 457-472.

Gelman, A., Van Dyk, D.A., Huang, Z. and Boscardin, W.J. (2008). Using redundant parameterizations to fit hierarchical models. *Journal of Computational and Graphical Statistics*, 17(1), 95-122.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A. and Rubin, D.B. (2004). *Bayesian Data Analysis*. New York: Chapman and Hall.

Geman, S., and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattn Anal. Mach. Intell.*, 6, 721-741.

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

Gonzalez, M.E., and Waksberg, J. (1973). Estimation of the error of synthetic estimates. Technical report, paper presented at the first meeting of the International Association of Survey Statisticians, Vienna, 18-25 August, 1973.

Hawala, S., and Lahiri, P. (2018). Variance modelling for domains. *Statisitcs and Applications*, 16, 399-409.

Hirose, M., and Lahiri, P. (2018). Estimating variance of random effects to solve multiple problems simultaneously. *Annals of Statistics*, 46, 1721-1741.

Hodges, J.S., and Sargent, D.J. (2001). Counting degrees of freedom in hierarchical and other richly-parameterised models. *Biometrika*, 88(2), 367-379.

Lahiri, P., and Pramanik, S. (2019). Evaluation of synthetic small-area estimators using design-based methods. *Austrian Journal of Statistics*, 48, 43-57.

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

Marhuenda, Y., Molina, I. and Morales, D. (2013). Small area estimation with spatio-temporal Fay-Herriot models. *Computational Statistics and Data Analysis*, 57, 308-325.

Marker, D. (1995). *Small Area Estimation: A Bayesian Perspective*. Ph.D. thesis, University of Michigan.

O'Malley, A.J., and Zaslavsky, A.M. (2008). Domain-level covariance analysis for multilevel survey data with structured nonresponse. *Journal of the American Statistical Association*, 103(484), 1405-1418.

Pfeffermann, D., and Ben-Hur, D. (2018). Estimation of randomisation mean square error in small area estimation. *International Statistical Review*, 87, 31-49.

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

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

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

Rao, J.N.K., and Molina, I. (2015). *Small Area Estimation*. Wiley-Interscience.

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

Rao, J.N.K., Rubin-Bleuer, S. and Estevao, V.M. (2018). Measuring uncertainty associated with model-based small area estimators. Unpublished research paper.

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

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

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B*, 64(4), 583-639.

Sugasawa, S., Tamae, H. and Kubokawa, T. (2017). Bayesian estimators for small area models shrinking both means and variances. *Scandinavian Journal of Statistics*, 44(1), 150-167.

van den Brakel, J.A. (2008). Design-based analysis of embedded experiments with applications in the Dutch Labour Force Survey. *Journal of the Royal Statisitcal Society*, 171, 581-613.

van den Brakel, J.A., and Boonstra, H.J. (2018). Hierarchical Bayesian bivariate Fay-Herriot model for estimating domain discontinuities. Discussion paper, Statistics Netherlands.

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

van den Brakel, J.A., Buelens, B. and Boonstra, H.J. (2016). Small area estimation to quantify discontinuities in repeated sample surveys. *Journal of the Royal Statisitcal Society*, 179, 229-250.

van den Brakel, J.A., Smith, P.A. and Compton, S. (2008). Quality procedures for survey transitions-experiments, time series and discontinuities. *Survey Research Methods*, 2, 123-141.

Vehtari, A., Gelman, A. and Gabry, J. (2015). `loo` : *Efficient Leave-one-out Cross-validation and WAIC for Bayesian Models*. R package version 0.1.3.

Vehtari, A., Gelman, A. and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistical Computation*, 27, 1413-1432.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11, 3571-3594.

Watanabe, S. (2013). A widely applicable Bayesian information criterion. *Journal of Machine Learning Research*, 14, 867-897.

Wolter, K. (2007). *Introduction to Variance Estimation*. Springer.

Ybarra, L.M.R., and Lohr, S.L. (2008). Small area estimation when auxiliary information is measured with error. *Biometrika*, 95(4), 919-931.

You, Y. (2008). Small area estimation using area level models with model checking and applications. *Proceedings of the Survey Methods Section of the Statistical Society of Canada*.

You, Y., and Chapman, B. (2006). Small area estimation using area level models and estimated sampling variances. *Survey Methodology*, 32, 1, 97-103. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2006001/article/9263-eng.pdf.

You, Y., and Rao, J.N.K. (2000). Hierarchical Bayes estimation of small area means using multi-level models. *Survey Methodology*, 26, 2, 173-181. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2000002/article/5537-eng.pdf.

Survey Methodology, June 2021
Vol. 47, No. 1, pp. 191-213
Statistics Canada, Catalogue No. 12-001-X

Bayesian pooling for analyzing categorical data from small areas

Aejeong Jo, Balgobin Nandram and Dal Ho Kim[1]

Abstract

Bayesian pooling strategies are used to solve precision problems related to statistical analyses of data from small areas. In such cases, the subpopulation samples are usually small, even though the population might not be. As an alternative, similar data can be pooled in order to reduce the number of parameters in the model. Many surveys consist of categorical data on each area, collected into a contingency table. We consider hierarchical Bayesian pooling models with a Dirichlet process prior for analyzing categorical data based on small areas. However, the prior used to pool such data frequently results in an overshrinkage problem. To mitigate for this problem, the parameters are separated into global and local effects. This study focuses on data pooling using a Dirichlet process prior. We compare the pooling models using bone mineral density (BMD) data taken from the Third National Health and Nutrition Examination Survey for the period 1988 to 1994 in the United States. Our analyses of the BMD data are performed using a Gibbs sampler and slice sampling to carry out the posterior computations.

Key Words: Categorical data; Dirichlet process; Nonparametric hierarchical Bayesian pooling; Slice sampling; Small area.

1. Introduction

Many surveys collect categorical data for individual areas, which are then stored in a contingency table. For example, in a typical obesity rate comparison survey, the researcher might classify the measured sample data by the degree of obesity. Then, the regional obesity rate is estimated using the number of samples assigned to each category. In such cases, we need to consider how the precision is affected by the sample size and the number of parameters in the model, particularly for estimations based on small areas (Rao and Molina, 2015). In general, the precision of a model decreases as the number of parameters increases, assuming the same sample data. To prevent this decrease in precision, the constructed model needs to be as simple as possible. That is, the number of parameters must be reduced in the model. However, the model loses the ability to reflect the detailed effects in each area. Another way to resolve the precision problem is to increase the sample size allowed per parameter. That is, we can employ pooling strategies when analyzing categorical data based on small areas.

Interest in pooling methods is growing among researchers. Malec and Sedransk (1992) developed a Bayesian procedure for estimating the mean of an experiment in a set of seemingly similar experiments. They constructed the prior distribution for a location parameter to reflect their assumptions. They identified subsets of parameters, with subscripts indicating the similarity between the subsets, in which there is uncertainty about the composition of the subsets. They specified the prior distribution for a parameter by conditioning on the same subscript in similar experiments. Their flexible prior distribution

1. Aejeong Jo, Researcher, Division for Healthcare Technology Assessment Research, National Evidence-based Healthcare Collaborating Agency, 173, Toegye-ro, Jung-gu, Seoul, 100-705, Korea. E-mail: joaejung@neca.re.kr; Balgobin Nandram, Professor, Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609, U.S.A. E-mail: balnan@wpi.edu; Dal Ho Kim, Professor, Department of Statistics, Kyungpook National University, 80 Daehakro, Bukgu, Daegu, 702-701, Korea. E-mail: dalkim@knu.ac.kr.

allows the intensity and nature of the pooling to be influenced by the sample data. Later, Evans and Sedransk (1999) proposed a more flexible Bayesian model using covariates. In addition, Evans and Sedransk (2003) provided a fully Bayesian justification for the results of Malec and Sedransk (1992). These three works have since been extended based on the same key concept of specifying a model in which subscripts are used to indicate similar experiments (Consonni and Veronese, 1995 and DuMouchel and Harris, 1983). Also, Dunson (2009) suggested a generalization of the Dirichlet process (DP) proposed by Ferguson (1973) that allows for dependent local pooling and the borrowing of information. The goal is to borrow information in order to more efficiently estimate the individual functions. The proposed process for local pooling offers a simple, but flexible approach to specifying the local selection process. They suggest using slice sampling, proposed by Walker (2007), to carry out the posterior computations. This is a simple and efficient method that allows for posterior computations for an infinite-dimensional process that is similar to those of a finite-dimensional process. Here, we construct a pooling model using these basic concepts for data based on small areas. Recently, Nandram, Zhou and Kim (2019) proposed a pooled Bayes test of independence for sparse contingency tables. They constructed the model based on a Dirichlet-multinomial hierarchical Bayesian model, see also Nandram (1998) who constructed a prior using the Dirichlet distribution for pooling the data in the models. Of course, a DP is assumed for the parameters of interest, which is the cell probability parameters in our contingency table.

In this study, we use bone mineral density (BMD) data, taken from the Third National Health and Nutrition Examination Survey (NHANES III) for the six-year period from October 1988 to September 1994. BMD is the quantity of mineral in bone tissue, measured as the optical density per cm^2 of bone surface using medical imaging. BMD is used in clinical arenas as an indirect indicator of osteopenia, osteoporosis, fractures, and so on. BMD is statistically correlated with the probability of fractures, which are an important public health problem, especially in elderly women. Therefore, BMD data are important indicators used to identify osteoporotic patients who might benefit from early management to improve their bone strength.

NHANES III contains clinical data on 33,994 people who participated in the survey and is sampled for individual areas. Each person is categorized into three BMD levels: (1) normal, (2) osteopenia and (3) osteoporosis. Our study used Bayesian inference on categorical tables. See Agresti and Hitchcock (2005) and Leonard (1977) for inference on second multinomial tables. The original data were gathered from mobile examination centers across the United States. NHANES III, which is an important program of the National Center for Health Statistics (NCHS), examines the state of health and nutritional in the United States. The program started in the 1960s, and has conducted surveys on various health- and nutrition-related topics. As a result, NHANES provides surveys based on large samples in the United States. However, the NCHS is also interested in estimates for smaller geographical areas and study domains. When the sample size of a subpopulation is small, we need to consider an alternative estimator based on a pooling strategy in order to analyze the data.

As a result, we focus on predicting the finite population proportion of each area. The finite population proportion is estimated by inputting the sample data into the model to predict the unobserved nonsample part of the finite population, then obtaining the weighted sum of the observed sample data and the predicted nonsample obtain to the sample proportion. First, we estimate the cell probability parameters

from the sample data. During this process, the observed count by category in NHANES III is employed as the sample part of the finite population. Second, these parameters are used to predict values for the nonsample part. Finally, we get the finite population proportions by combining the sample data and prediction value of nonsample part.

The remainder of the paper proceeds as follows. In Section 2, we introduce the hierarchical Bayesian pooling strategies used to analyze categorical data from small areas. In Section 3, we present and discuss the results of our data analysis of the BMD survey data set. Section 4 concludes the paper. Appendix A and B include the computation process for hierarchical Bayesian pooling model.

2. Hierarchical Bayesian models

2.1 Parametric models

For a hierarchical Bayesian baseline model, we consider an $I \times K$ contingency table, where n_{ik} indicates the k^{th} response in the i^{th} area, for $i = 1, \ldots, I$, $k = 1, \ldots, K$. Let π_{ik} denote the corresponding proportion for each cell. Then, we assume that

$$\mathbf{n}_i \,|\, \boldsymbol{\pi}_i \overset{\text{iid}}{\sim} \text{Multinominal}(n_{i.}, \boldsymbol{\pi}_i), \tag{2.1}$$

where $\mathbf{n}_i = (n_{i1}, \ldots, n_{iK})$ for $i = 1, \ldots, I$, is a vector of responses, $n_{i.} = \sum_{k=1}^{K} n_{ik}$ is the sum of the responses in area i, and $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iK})$, $0 \le \pi_{ik} \le 1$, $\sum_{k=1}^{K} \pi_{ik} = 1$, is the proportion vector for each area. The model does not allow any pooling and is denoted as a baseline to compare our models. The parameters π_i are independent and do not share a common effect. That is, the areas are unrelated.

There are five categories of parametric pooling models, classified according to the priors of the proportion vectors in a multinomial distribution. First, the four prior distributions for parametric Bayesian inferences are given as follows:

1) No pooling, $\boldsymbol{\pi}_i \overset{\text{iid}}{\sim} \text{Dirichlet}(\mathbf{1})$;
2) Complete pooling, $\boldsymbol{\pi} \sim \text{Dirichlet}(\mathbf{1})$ with $\boldsymbol{\pi}_1 = \ldots = \boldsymbol{\pi}_I = \boldsymbol{\pi}$;
3) Adaptive pooling, $\boldsymbol{\pi}_i \overset{\text{iid}}{\sim} \text{Dirichlet}(\boldsymbol{\mu}\tau)$;
4) Restricted pooling, $\boldsymbol{\pi}_i \overset{\text{ind}}{\sim} \phi\ \text{Dirichlet}(\boldsymbol{\mu}\tau) + (1 - \phi)\,\text{Dirichlet}(\mathbf{1})$;

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)'$, $0 \le \mu_k \le 1$, $\sum_{k=1}^{K} \mu_k = 1$ and $\tau > 0$ are the hyperparameters of the Dirichlet distribution. We further assume that $\pi(\boldsymbol{\mu}, \tau) = (K - 1)!/(1 + \tau)^2$, a shrinkage prior. We note that Yin and Nandram (2020a, b) place a Dirichlet process on the sampling process to accomodate gaps, outliers and ties in survey data, see also Nandram and Yin (2016a, b) for additional discussion of the Dirichlet process. The $\phi \sim \text{uniform}(1/2, 1)$, $\phi > 1/2$ means that more weight is attached to adaptive pooling.

Model 1 is a no-pooling model that estimates the parameter without any data sharing from other areas. Model 2, on the other hand, is a complete pooling model that estimates the parameter while treating the different areas as one. When conducting parameter estimations on a small area with a small number of data using Model 1, the estimation may face the small area problem, as the parameter is estimated by relying on insufficient data. Although the complete pooling model alleviates this issue, it faces problems

of its own: Overshrinkage and individual areas cannot be discerned. Hence, this paper introduces various pooling approaches to find a model that delivers better estimates. In Model 3, the adaptive pooling model introduced by Nandram, Zhou and Kim (2019), all areas share the same hyperparameters; hence, they share their area data information as well. This is an indirect complete pooling method that preserves some area variation but, in general, assumes all areas to have identical traits; see also Nandram (1998). This creates variations in estimating the hyperparameters. Thus, we propose the restricted pooling model, which alters Model 3 by removing data-sharing information between distinct areas. In this new model, distinct areas use their local data to estimate parameters, and areas with similar traits share information through pooling based on the same hyperparameter, thereby improving the estimation. This model, however, assigns the same hyperparameter to areas with similar traits, which may lead to the overshrinkage problem when smoothing in the category occurs. To mitigate this issue, we propose Model 5, the global-local pooling model, see Dunson (2009). The global-local pooling model pools information in data among areas with similar traits but also preserves each variation in the category through the local effect model, thereby reducing the smoothing in the categorical effect. Indeed, Model 5 is flexible and robust.

The fifth prior distribution used for parametric Bayesian inferences is called global–local pooling. In this case, we use different notation for the proportion vector of each area. Let $\mathbf{p}_i$, for $i = 1, \ldots, I$, denote the corresponding cell proportion vector in the i^{th} area. We assume that

$$\mathbf{n}_i \,|\, \mathbf{p}_i \overset{\text{ind}}{\sim} \text{Multinomial}(n_{i.}, \mathbf{p}_i),\ i = 1, \ldots, I, \tag{2.2}$$

where $\mathbf{p}_i = \left(e^{\theta+\eta_{i1}} \Big/ \left(1 + \sum_{k=1}^{K-1} e^{\theta+\eta_{ik}}\right), \ldots, e^{\theta+\eta_{i(K-1)}} \Big/ \left(1 + \sum_{k=1}^{K-1} e^{\theta+\eta_{ik}}\right), 1 \Big/ \left(1 + \sum_{k=1}^{K-1} e^{\theta+\eta_{ik}}\right)\right)'$. Here, $\mathbf{p}_i$ is composed of two components, namely, θ and η_{ik}, and θ reflects the basic probability that brought all the areas together. The global-local effect is reflected in the component $\boldsymbol{\eta}_i = \left(\eta_{i1}, \ldots, \eta_{i(K-1)}\right)'$. Specifically,

$$\boldsymbol{\eta}_i \overset{\text{iid}}{\sim} I_{(z_i=0)} \prod_{k=1}^{K-1} \text{N}(0, \sigma^2) + I_{(z_i=1)} \prod_{k=1}^{K-1} \text{N}(0, \sigma_k^2), \tag{2.3}$$

where $\text{N}(0, \sigma^2)$ is the normal distribution of the global parameter σ^2, $\text{N}(0, \sigma_k^2)$ is the normal distribution of the local parameters σ_k^2 for each category, where each area is denoted by a different index. Then, $z_i, i = 1, \ldots, I,$ follows a Bernoulli distribution with a hyperparameter ϕ, which adjusts the proportion between the global and local effects. Thus, if we need to focus on the global effect, the prior for ϕ is set using the uniform distribution on $(1/2, 1)$. Specifically, we assume that

$$z_i \,|\, \phi \overset{\text{iid}}{\sim} \text{Bernoulli}(\phi),\ i = 1, \ldots, I,$$

$$\phi \sim \text{Uniform}\left(\frac{1}{2}, 1\right),$$

$$\pi(\theta) = \frac{1}{\pi(1 + \theta^2)}, \text{ a Cauchy prior,}$$

$$\pi(\sigma^2, \sigma_1^2, \ldots, \sigma_{K-1}^2) = \frac{1}{(1 + \sigma^2)^2} \prod_{k=1}^{K-1} \frac{1}{(1 + \sigma_k^2)^2},$$

where $-\infty < \eta_{ik} < \infty$, $k = 1, \ldots, K-1$, $\sigma^2 > 0$, and $\sigma_k^2 > 0$, $k = 1, \ldots, K-1$. We have used the Cauchy prior for θ and shrinkage priors for variance components. The shrinkage priors are similar to the half Cauchy prior and they are mathematically more convenient when we make transformations to $(0, 1)$.

The models proposed in this paper are based on the adaptive pooling model (Nandram, Zhou and Kim, 2019), which applies the principle of assigning the same subscripts of parameter in prior distribution to similar experiments (Malec and Sedransk, 1992) to categorical data. In particular, the no pooling model and complete pooling model represent two extreme cases of adaptive pooling, with parameters $\boldsymbol{\mu}$ and τ. On the other hand, the restricted pooling model has a pooling principle such as adaptive pooling model, but the model also reflects uncertainty through the weighting parameter ϕ. However, the same parameter in the prior distribution used to pool such data frequently results in an overshrinkage problem. To compensate for this problem, we propose the global-local pooling model. The effects of the parameters are separated into global and local effects in this model. As a result, we propose two new models, restricted pooling model and global-local pooling model, and we compare these models with existing ones.

2.2 Nonparametric models

We consider the DP prior for $\boldsymbol{\pi}_i$ of (2.1) in Section 2.1. The prior structure is as follows:

$$\pi_i \,|\, G \overset{\text{iid}}{\sim} G$$
$$G \sim \text{DP}(\alpha, G_0),$$

where G_0 is the base distribution and the positive real number, α, is the concentration parameter in the DP prior. The model is specified by the structure of the base distribution. We note here that Yin and Nandram (2020a, b) used a DP on the sampling process to accommodate gaps, outliers and ties in survey data. First, we define the model using two prior distributions, as follows:

6) Nonparametric adaptive pooling $\quad G_0 \equiv \text{Dirichlet}(\boldsymbol{\mu}\tau)$;

7) Nonparametric restricted pooling $\quad G_0 \equiv \phi\ \text{Dirichlet}(\boldsymbol{\mu}\tau) + (1-\phi)\,\text{Dirichlet}\left(\mathbf{1}_{1\times K}\right)$.

We assume $\pi(\boldsymbol{\mu}, \tau) = (K-1)!/(1+\tau)^2$, $\pi(\alpha) \propto 1/(1+\alpha)^2$, and $\phi \sim \text{uniform}(1/2, 1)$. In Models 6 and 7, we use a stict-breaking process for the DP prior (Sethuraman, 1994).

The last model is a nonparametric version of (2.2) in Section 2.1, used for global-local pooling. Here, $\boldsymbol{\eta}_i$ are used to construct the nonparametric Bayesian setting, as follows:

$$\boldsymbol{\eta}_i \overset{\text{iid}}{\sim} I_{(z_i=0)}G_0 + I_{(z_i=1)}\prod_{k=1}^{K-1} G_k$$

$$G_0 \sim \text{DP}(\alpha_0, \text{MVN}(\mathbf{0}, \sigma^2\mathbf{I})),\ G_k \sim \text{DP}(\alpha_1, N(0, \sigma_k^2)),$$

where $\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{i(K-1)})'$, $z_i \sim \text{Bernouilli}(\phi)$, $\pi(\theta) \propto 1$, $\phi \sim \text{uniform}(1/2, 1)$, $\pi(\sigma^2, \sigma_1^2, \ldots, \sigma_{K-1}^2) = 1\Big/\left\{(1+\sigma^2)^2 \prod_{k=1}^{K-1}(1+\sigma_k^2)^2\right\}$, for $i = 1, \ldots, I$. This is model (A.1), nonparametric global-local pooling.

The distribution of η_i involves a mixture of global and local pooling areas. While global pooling is conducted according to the same principle as the aforementioned nonparametric models, the Dirichlet

process prior, where the normal distribution is the base distribution for each cell, is independently defined, thereby alleviating the overshrinkage problem that could arise owing to global pooling. Here, z_i, which represents the weight of the model, follows the Bernouilli distribution with ϕ greater than $1/2$ as its parameter, and ϕ follows the uniform distribution. Hence, the local pooling area is weighted so that it is defined as the form that can better alleviate the degree of shrinkage. In addition, to ensure simplicity of the model, the heavy tail and noninformative characteristic of θ distribution follow the improper prior which is identical to, but simpler than, the previous parametric model and the posterior distribution is presented properly.

3. Data analysis

In this section, we present the empirical results from comparing the performance of the five parametric pooling models described in Section 2 and the three nonparametric pooling versions described in Section 3. We use BMD data for the period 1988 to 1994, taken from the NHANES III, which collected data from mobile examination centers across the United States.

Our analysis is conducted using contingency tables with a cell count for three categories of BMD in 31 counties in the U.S. Here, BMD is categorized into one of three levels. The normal category is defined as those with a BMD value less than one standard deviation (SD) below the non-Hispanic white (NHW) adult mean $(0.82\,\mathrm{mg/cm^2} < \mathrm{BMD})$. The osteopenia category is defined as a BMD value between 1 and 2.5 SD below the young NHW adult mean $(0.64\,\mathrm{mg/cm^2} < \mathrm{BMD} \le 0.82\,\mathrm{mg/cm^2})$. Then, the osteoporosis category corresponds to a BMD of more than 2.5 SD below the young NHW adult mean $(\mathrm{BMD} \le 0.64\,\mathrm{mg/cm^2})$.

We predict the finite population proportion for the BMD distribution in each area using the Bayesian pooling model. The survey covers roughly 0.02% of the population, and prediction as needed for the remainder 99.98%, an enormous job. Table 3.1 shows the sample data, which have a cell count for each categorized level in each area. We estimate the finite population proportion by predicting the nonsample part of the finite population from a multinomial distribution with parameter $\boldsymbol{\pi}_i$, $i = 1, \ldots, I\,(= 31)$ at each MCMC iteration. Specifically, let N_{ik} for $k = 1, 2, 3$ be the total BMD level in area i, where the value is unknown. We have the value (n_{ik}) for the sample part of the finite population. Then, we compute the finite population proportion (P_{ik}) for $i = 1, \ldots, I\,(I = 31)$ as follows:

$$P_{ik} = \frac{1}{N_i}\{n_{ik} + (N_{ik} - n_{ik})\},\ k = 1, 2, 3, \tag{3.1}$$

where $N_i = \sum_{k=1}^{3} N_{ik}$, $N_{ik} - n_{ik}$ is the nonsample part for each BMD level $k\,(k = 1, 2, 3)$ in area i, taken from the multinomial distribution with parameter $\hat{\boldsymbol{\pi}}_i$, estimated using the MCMC in each model. Then, the posterior mean and standard deviation of P_{ik} are obtained using the estimated empirical distribution of P_{ik}.

Table 3.1
BMD categorical data for 31 counties from NHANES III

Areas	BMD		
	Normal	Osteopenia	Osteoporosis
1	33	24	9
2	46	39	5
3	40	25	8
4	48	25	6
5	40	15	10
6	74	30	12
7	47	19	7
8	38	15	6
9	49	16	11
10	99	40	14
11	39	18	2
12	63	27	4
13	48	18	5
14	42	16	4
15	40	15	4
16	110	44	7
17	37	14	3
18	55	18	5
19	47	12	6
20	296	95	17
21	59	18	4
22	78	21	7
23	196	55	15
24	149	44	9
25	69	19	5
26	49	10	6
27	73	19	3
28	76	14	3
29	77	13	4
30	96	13	6
31	88	12	4

We use 1,000 iterations to "burn in" the MCMC samples, and take every 10^{th} value to obtain the 1,000 iterations. In addition, we use autocorrelation plots of the model to adjust the number of repetitions and thinning intervals. For example, in a nonparametric model with a relatively large number of parameters, we take every 20^{th} estimated value from 1,001 to 20,000. We set the initial value of proportion $\boldsymbol{\pi}_i$ for $i = 1, \ldots, 31$ based on the column proportion of the sample values in each area.

The groups are categorized according to the quartile values of the first column proportion. The tuning parameter ξ_j, $j = 1, \ldots, J$, is initially set to $\xi = 0.5$, and then is revised based on the performance of each model.

In Table 3.2, we report the posterior means (PM) and posterior standard deviations (PSD) of the finite population proportions for the eight models and some areas. The cases of Model 1 and Model 2 are the most extreme pooling structures. The PM of Model 1 has the results 0.511, 0.496, 0.901, 0.826, and 0.820 for the corresponding areas 1, 2, 28, 29, and 31, respectively, in the normal BMD, implying that the areas' fluctuations are greater than those in Model 2's PM (0.652, 0.654, 0.714, 0.716, 0.719). For Model 3, the fluctuations (0.644, 0.612, 0.793, 0.816, 0.798) show a trend similar to that in Model 1's PM for each area but are smoother than those in Model 1. This could be interpreted as the indirect pooling effect through the

hyper-parameter rather than the direct pooling effect in Model 2. For the restricted pooling, in areas that show similar characteristics, the estimated values are calculated through indirect pooling and hyper-parameters as in Model 3, and the estimated values are smoother than those in Model 1 (PM of areas 28, 29, 31: 0.827, 0.779, 0.864, respectively). However, in areas where similar characteristics are not shown, the estimated values are close to those in Model 1 because the parameter is estimated by solely relying on the information in its associated areas (PM of areas 1, 2: 0.523, 0.510, respectively). Model 5, which is proposed to alleviate the overshrinkage problem that could arise owing to information pooling, shares information in nearby areas and alleviates the excessive shrinkage by reflecting the local effect of each cell, thereby rendering the estimated values that are between those in Model 1 and Model 2 (PM of Model 5: 0.581, 0.511, 0.711, 0.836, 0.808). It should be noted that the nonparametric Bayesian model assigns the areas with indexes as the same group according to the characteristics of information through hyper-parameter d_i, $i = 1, \ldots, I\,(= 31)$, and the same group shares the parameter directly. For BMD data, the number of the group ranges from one to three, showing the highest frequency, and the PM is estimated as in Model 2. The characteristics of the estimated values of PM for each model are shown identically in osteopenia BMD and osteoporosis BMD as well.

The posterior means, standard deviations (SD) and posterior coefficients of variation (CV) of the finite population proportions for the eight models can also be seen in Figure 3.1-3.3.

In Figure 3.1, the variation of PM for eight models is the largest in the normal BMD, which takes up the largest proportion. Especially, we can see that PMs of the nonparametric model are similar to that of the complete pooling model. During data analysis, we found one to three group index. Through this, we were able to discover that the BMD distribution is quite similar across the areas in NHANES III. Hence, all areas share their information with others to estimate the same hyperparameter. Thus, the nonparametric and complete pooling models give similar estimates. Therefore, pooling can solve problems associated with small area estimates when areas share similar traits. Furthermore, Figure 3.2 shows that the performance of the nonparametric models are good through the fact that the SD of the nonparametric models with many parameters are similar or smaller than that of the parametric models.

In addition, we can see the CV of the models by BMD status in Figure 3.3. In the case of CV, osteoporosis BMD shows the greatest difference between models. Also it can be seen that the CV of the nonparametric versions is relatively low compared to the parametric version, which is not different for each BMD status. Furthermore, it is very meaningful that the nonparametric version had a smaller CV than the models of the parametric version, even though it had infinite parameter space.

To estimate the parameters, we use a Gibbs sampler. Whereas the parameters with restricted parameter spaces are sampled using the grid method, the other parameters are sampled using the Metropolis-Hastings algorithm. We tune to get acceptance rate 30-70%. In the actual analysis, the acceptance rate of the algorithm is 34-49%. We compare the two measures in terms of the performance of each model. First, we calculate the deviance information criterion (DIC), a typical Bayesian model choice criterion, to compare the hierarchical Bayesian models. The DIC was proposed by Spiegelhalter et al. (2002), where a lower DIC value indicates better performance. Second, we evaluate the performance of the eight models by

calculating the logarithmic conditional predictive ordinate (LCPO), which is a comparison method that uses cross validation. The average LCPO, proposed by Gneiting and Raftery (2007), is calculated as follows:

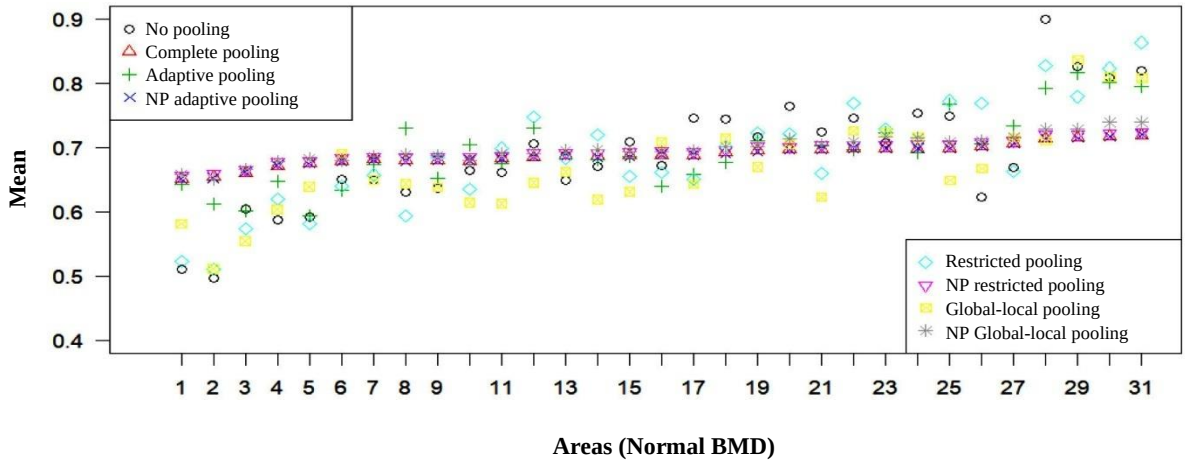
$$\overline{\text{LCPO}} = -\frac{1}{I}\sum_{i=1}^{I}\log\left(\hat{\text{CPO}}_i\right), \tag{3.2}$$

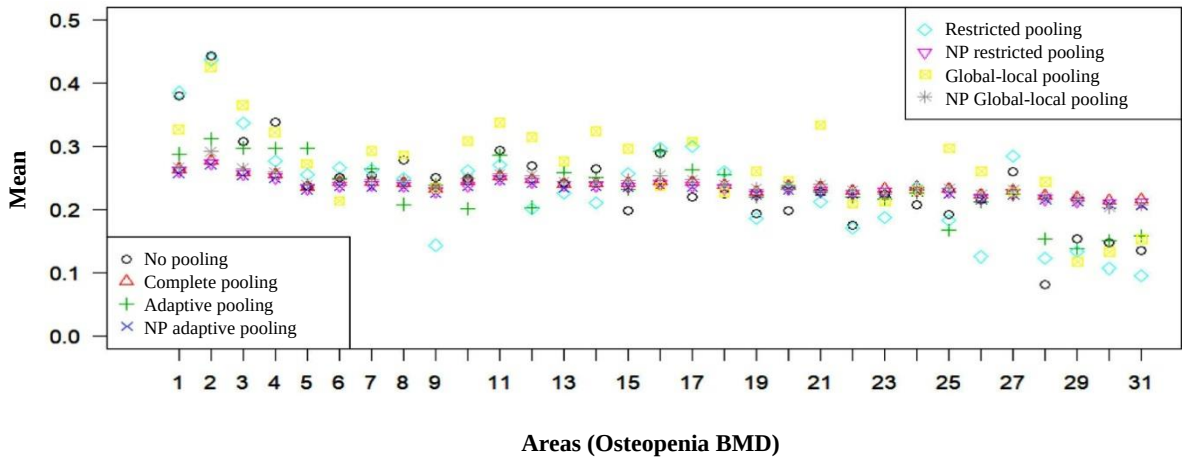
where $\hat{\text{CPO}}_i = \sum_{h=1}^{H} w_h P(Y = y|\boldsymbol{\Omega}^{(h)})$, $w_h = \sum_{h=1}^{H} f(Y = y|\boldsymbol{\Omega}^{(h)}) / f(Y = y|\boldsymbol{\Omega}^{(h)})$, for $i = 1, \ldots, I$, and $P(Y = y|\boldsymbol{\Omega}^{(h)})$ is the likelihood of a single observation of a given parameter $\boldsymbol{\Omega}^{(h)}$, and $h = 1, \ldots, H$ denotes the iterations from the MCMC result under the hierarchical Bayesian pooling model.

Table 3.2
Posterior summaries for finite population proportions of BMD data under the eight models by areas

Areas	Model	Normal BMD			Osteopenia BMD			Osteoporosis BMD		
		PM	PSD	95% CI	PM	PSD	95% CI	PM	PSD	95% CI
1	1	0.511	0.024	(0.467, 0.561)	0.380	0.024	(0.333, 0.427)	0.109	0.015	(0.082, 0.139)
	2	0.652	0.023	(0.603, 0.692)	0.264	0.021	(0.227, 0.306)	0.084	0.013	(0.061, 0.109)
	3	0.644	0.024	(0.594, 0.691)	0.288	0.023	(0.242, 0.338)	0.068	0.011	(0.048, 0.091)
	4	0.523	0.024	(0.476, 0.570)	0.385	0.024	(0.336, 0.432)	0.092	0.013	(0.070, 0.118)
	5	0.581	0.022	(0.536, 0.627)	0.327	0.022	(0.285, 0.370)	0.092	0.013	(0.067, 0.118)
	6	0.651	0.022	(0.606, 0.696)	0.258	0.020	(0.218, 0.297)	0.090	0.014	(0.067, 0.117)
	7	0.656	0.023	(0.612, 0.700)	0.260	0.021	(0.218, 0.300)	0.084	0.013	(0.061, 0.109)
	8	0.658	0.022	(0.615, 0.703)	0.267	0.021	(0.230, 0.311)	0.075	0.012	(0.052, 0.100)
2	1	0.496	0.021	(0.458, 0.534)	0.442	0.021	(0.402, 0.484)	0.061	0.010	(0.042, 0.080)
	2	0.654	0.021	(0.617, 0.693)	0.277	0.019	(0.242, 0.316)	0.068	0.010	(0.047, 0.087)
	3	0.612	0.020	(0.570, 0.652)	0.313	0.019	(0.273, 0.353)	0.075	0.012	(0.053, 0.098)
	4	0.510	0.022	(0.470, 0.551)	0.438	0.022	(0.396, 0.478)	0.052	0.009	(0.036, 0.071)
	5	0.511	0.023	(0.466, 0.554)	0.425	0.023	(0.381, 0.472)	0.064	0.011	(0.045, 0.085)
	6	0.653	0.020	(0.618, 0.693)	0.271	0.017	(0.237, 0.304)	0.075	0.012	(0.053, 0.096)
	7	0.659	0.020	(0.618, 0.702)	0.273	0.018	(0.239, 0.311)	0.068	0.011	(0.047, 0.089)
	8	0.651	0.021	(0.609, 0.689)	0.292	0.019	(0.254, 0.329)	0.058	0.011	(0.040, 0.079)
28	1	0.901	0.011	(0.881, 0.920)	0.081	0.010	(0.062, 0.099)	0.018	0.005	(0.010, 0.028)
	2	0.714	0.020	(0.677, 0.751)	0.223	0.017	(0.189, 0.256)	0.064	0.011	(0.043, 0.088)
	3	0.793	0.017	(0.757, 0.826)	0.154	0.015	(0.126, 0.184)	0.053	0.010	(0.034, 0.073)
	4	0.827	0.016	(0.796, 0.856)	0.123	0.013	(0.099, 0.148)	0.050	0.010	(0.032, 0.069)
	5	0.711	0.021	(0.672, 0.751)	0.244	0.020	(0.209, 0.283)	0.046	0.010	(0.028, 0.064)
	6	0.715	0.018	(0.680, 0.748)	0.216	0.016	(0.184, 0.249)	0.070	0.011	(0.049, 0.092)
	7	0.721	0.019	(0.683, 0.757)	0.217	0.018	(0.185, 0.251)	0.062	0.011	(0.043, 0.084)
	8	0.729	0.022	(0.686, 0.770)	0.220	0.020	(0.181, 0.261)	0.052	0.011	(0.034, 0.073)
29	1	0.826	0.015	(0.796, 0.855)	0.153	0.015	(0.124, 0.184)	0.020	0.005	(0.011, 0.030)
	2	0.716	0.019	(0.679, 0.755)	0.219	0.017	(0.187, 0.253)	0.065	0.010	(0.047, 0.085)
	3	0.816	0.016	(0.785, 0.847)	0.139	0.014	(0.113, 0.168)	0.045	0.008	(0.030, 0.060)
	4	0.779	0.017	(0.746, 0.815)	0.134	0.014	(0.105, 0.160)	0.087	0.012	(0.065, 0.111)
	5	0.836	0.016	(0.807, 0.866)	0.117	0.013	(0.095, 0.145)	0.047	0.009	(0.031, 0.067)
	6	0.715	0.020	(0.677, 0.755)	0.213	0.018	(0.179, 0.248)	0.072	0.012	(0.051, 0.096)
	7	0.721	0.019	(0.685, 0.756)	0.214	0.017	(0.183, 0.251)	0.066	0.011	(0.045, 0.085)
	8	0.729	0.021	(0.690, 0.768)	0.217	0.020	(0.179, 0.257)	0.053	0.011	(0.034, 0.075)
31	1	0.820	0.015	(0.792, 0.849)	0.136	0.014	(0.110, 0.161)	0.044	0.008	(0.030, 0.061)
	2	0.719	0.019	(0.680, 0.758)	0.216	0.017	(0.181, 0.253)	0.065	0.010	(0.046, 0.087)
	3	0.796	0.016	(0.765, 0.827)	0.159	0.015	(0.133, 0.188)	0.046	0.008	(0.031, 0.063)
	4	0.864	0.013	(0.838, 0.889)	0.095	0.011	(0.075, 0.119)	0.041	0.008	(0.027, 0.056)
	5	0.808	0.018	(0.773, 0.841)	0.153	0.017	(0.122, 0.188)	0.038	0.008	(0.024, 0.054)
	6	0.721	0.018	(0.685, 0.756)	0.207	0.017	(0.177, 0.239)	0.072	0.011	(0.052, 0.092)
	7	0.724	0.018	(0.688, 0.758)	0.211	0.017	(0.180, 0.243)	0.065	0.010	(0.045, 0.085)
	8	0.739	0.020	(0.698, 0.774)	0.208	0.019	(0.171, 0.247)	0.053	0.011	(0.033, 0.073)

Figure 3.1 The posterior means plot of the finite population proportion.





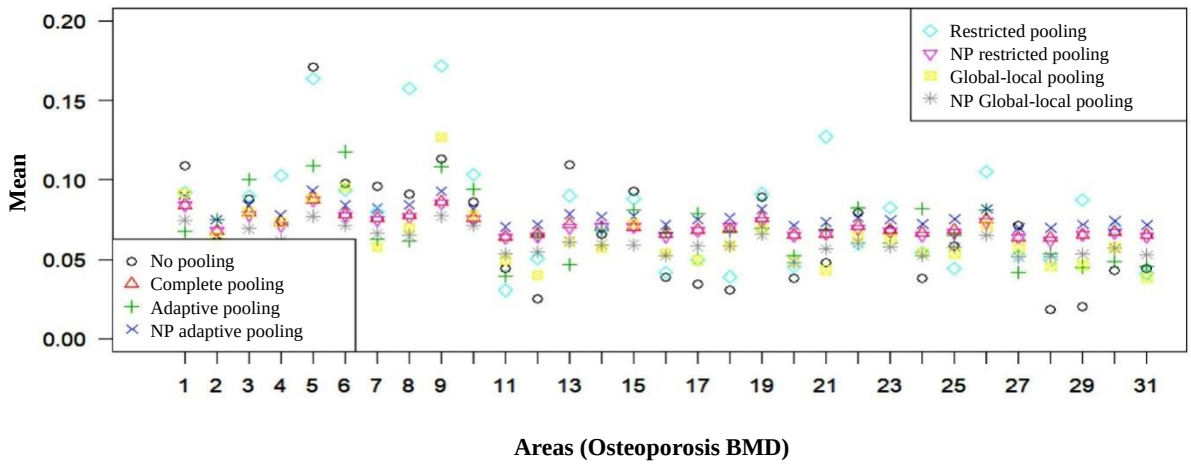
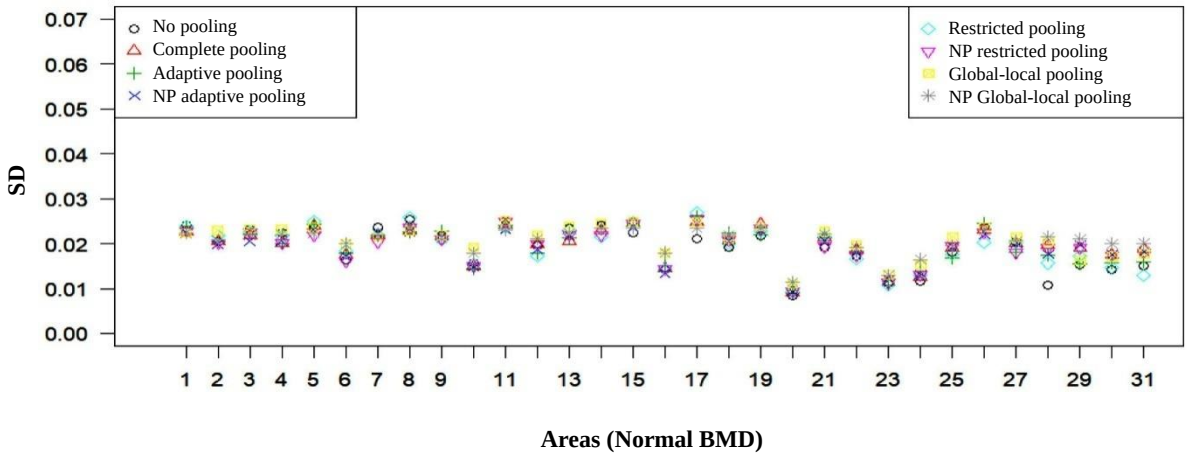
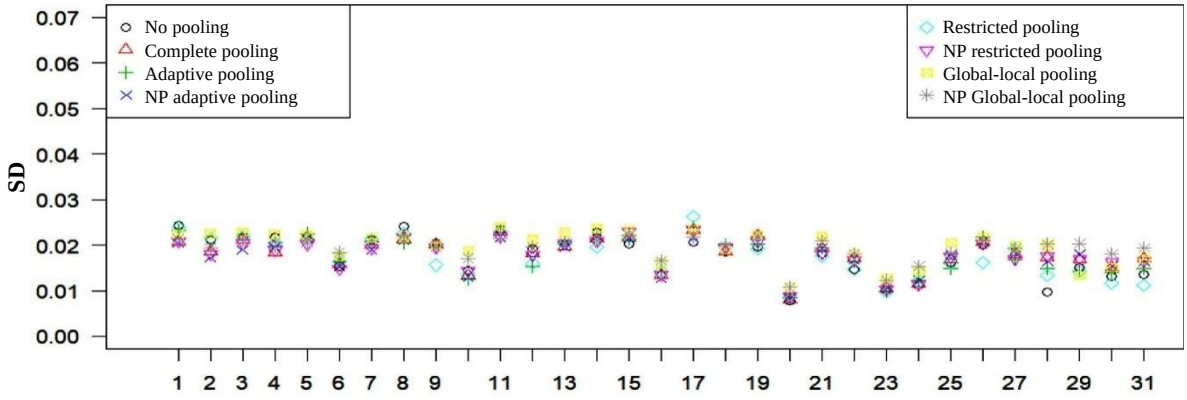


Figure 3.2 The posterior standard deviations plot of the finite population proportion.





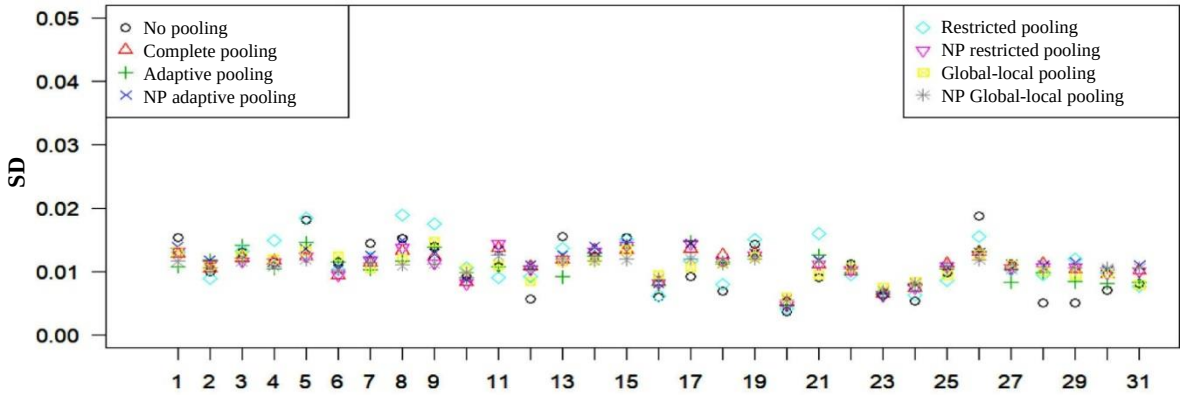
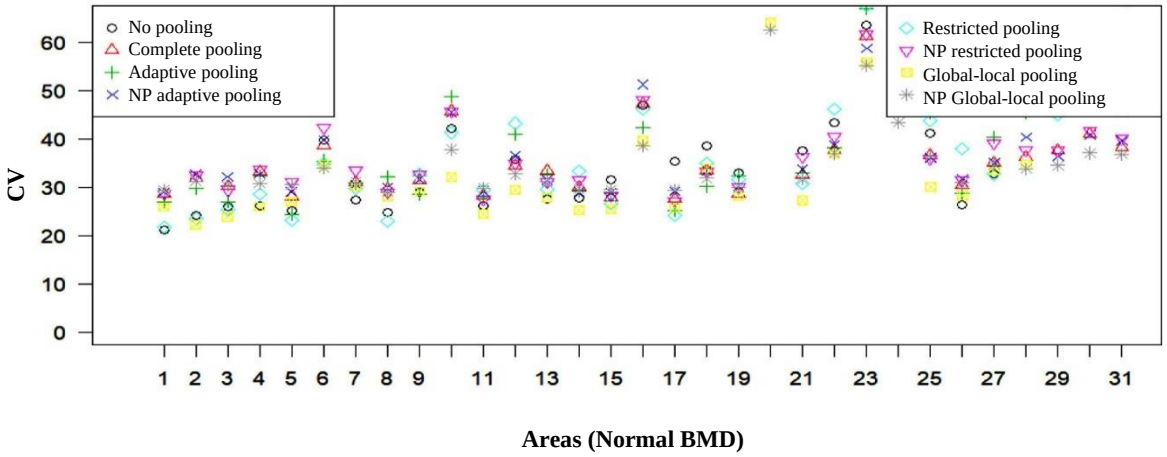
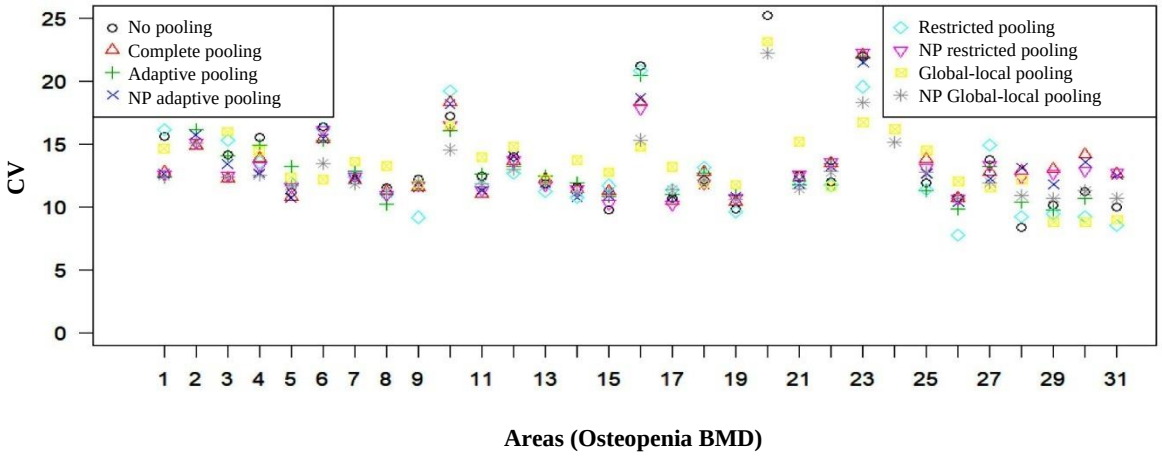


Figure 3.3 The coefficients of variation plot of the finite population proportion.





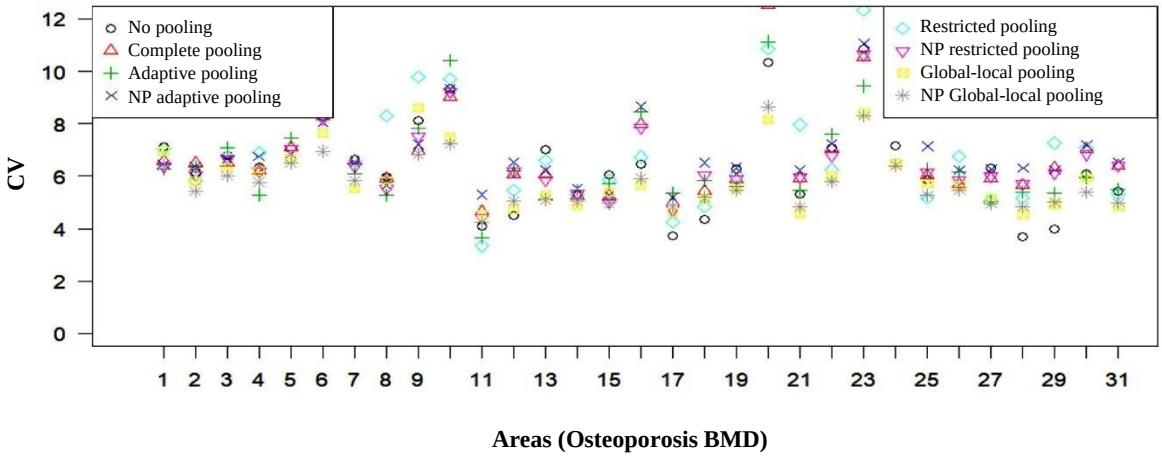


Table 3.3 shows the results of the two measures for each Bayesian pooling model. The model is considered to perform better as its estimated measures are smaller. The LCPO and DIC in parametric models are compared using the global-local pooling model, and LCPO and DIC have 12.707 and 4,984.56, respectively, implying the best performance. The restricted pooling model in Model 4 has an LCPO value of 12.886, showing the second-best performance after the global-local pooling model, but the DIC in the adaptive pooling model has a value of 4,998.90, which is lower than that in the restricted pooling model.

Table 3.3
Comparisons of $\overline{\text{LCPO}}$ and DIC (95% CI) under Models 1-8

Model		$\overline{\text{LCPO}}$	DIC
Parametric models	No pooling	13.167	4,999.19
	Complete pooling	13.151	5,011.45
	Adaptive pooling	13.411	4,998.90
	Restricted pooling	12.886	5,000.21
	Global-local pooling	12.707	4,984.56
Nonparametric models	Adaptive pooling	13.105	5,001.17
	Restricted pooling	12.837	4,983.88
	Global-local pooling	12.694	4,768.47

Another point to keep in mind is that the nonparametric version using the same pooling method shows similar values to those of the parametric method. In particular, although the nonparametric global-local pooling model has the greatest number of parameters to be estimated, its LCPO and DIC have values of 12.694 and 4,768.47, respectively, thereby indicating that it has the best performance among all the models. Additionally, in the restricted pooling model, the nonparametric model's LCPO and DIC have values of 12.837 and 4,983.88, respectively, showing better performance than the parametric model. These results are identical in the LCPO scale of the adaptive pooling (i.e., base) model (LCPO (parametric vs nonparametric) = (13.411 vs 13.105)). This means the performance of the nonparametric version is very good for our data, even though the parameter space has infinite dimensions.

Table 3.4 illustrates the calculated statistical values to estimate the shrinkage of the model. To estimate shrinkage, we calculated the average and standard deviation of the absolute difference from the no shrinkage model for each BMD category. Let PM_{ik}, $i = 1, \ldots, I$, $k = 1, 2, 3$ denote the posterior mean of the finite population proportion for each cell in area i. The average (ASE) and standard deviation (SDSE) of the shrinkage estimator are

$$\text{ASE}_k = \frac{1}{I}\sum_{i=1}^{I}\frac{|\text{PM}_{ik} - \text{PM}_{0ik}|}{\text{PM}_{0ik}},\ I = 31, \tag{3.3}$$

$$\text{SDSE}_k = \sqrt{\frac{1}{I-1}\sum_{i=1}^{I}\left(\frac{|\text{PM}_{ik} - \text{PM}_{0ik}|}{\text{PM}_{0ik}} - \text{ASE}_k\right)^2}, \tag{3.4}$$

where PM_{0ik} is the posterior mean for the finite population proportion of Model 1, a no poooling model.

Based on this calculation, we show that the ASE and SDSE in the normal cell are the smallest in the global-local pooling model. Drawing from the data analysis, with the BMD data applied in this study, the number of groups of slice sampling is 1 to 3, and we can confirm that the data characteristics are identical for most regions. In the global-local model, however, it can be shown that the problem of overshrinkage induced by pooling is solved by looking at the smallest shrinkage degree. In addition, osteopenia cells and ostoporosis cells could confirm the tendency of low shrinkage relative to other models, and the SDSE in the global-local pooling model can be shown to be small. Meanwhile, in the case of nonparametric models, the group index had the highest number of 1 because of slice sampling; therefore, it could be suspected that data dependence was excessive in the no pooling model.

Also, Geweke's test, autocorrelation plot and effective sample size (ESS) were applied for the diagnosis of the model, and they showed strongly mixing chains.

Table 3.4
A comparision of shrinkage in the eight models, see equations (3.3) and (3.4)

Model	Normal		Osteopenia		Osteoporosis	
	Mean	Std	Mean	Std	Mean	Std
No pooling(no shrinkage model)						
Complete pooling	0.087	0.072	0.217	0.320	0.522	0.622
Adaptive pooling	0.066	0.064	0.147	0.166	0.474	0.501
Restricted pooling	0.088	0.071	0.210	0.302	0.587	0.728
Global-local pooling	0.054	0.043	0.157	0.137	0.456	0.682
NP Adaptive pooling	0.088	0.074	0.209	0.306	0.524	0.609
NP Restricted pooling	0.065	0.049	0.249	0.354	0.335	0.362
NP Global-local pooling	0.085	0.072	0.206	0.311	0.427	0.430

NP: Nonparametric

4. Conclusion

In this study, we construct hierarchical parametric Bayesian pooling models and their nonparametric versions using the Ferguson (1973) Dirichlet process prior to pool the data. The pooling methodologies developed here are useful for analyzing survey data. We used the grid method to draw the parameters with nonstandard posterior densities and support that lies in a finite interval. However, we used the Metropolis-Hastings algorithm to draw the parameters with support in an infinite interval.

The Dirichlet process is assumed for the parameter of interest $\boldsymbol{\pi}_i$, for $i = 1, \ldots, I$, in our models. We apply the slice sampling algorithm for the specification of Dirichlet process prior, which is an extension of the widely used stick-breaking prior proposed by Ishwaran and James (2001). Five parametric models are modeled in a finite-dimensional parameter space, and three nonparametric versions have an infinite-dimensional parameter space. The eight hierarchical Bayesian models are also distinguished according to the type of effects in the model parameters. For the basic model (2.1), we can construct more effective and

efficient models that allow for a borrowing effect from neighboring areas in small-area estimations. However, exchangeable priors in a hierarchical Bayesian model may cause an overshrinkage problem. To compensate for this problem, the effect of a parameter is divided into two elements, as shown in the basic model (2.2), called the hierarchical Bayesian global-local pooling model. The model allows for grouping of similar experiments (area) and the borrowing of information in each area.

To compare the eight models using real data, we use the BMD data provided by the NHANES III. BMD is statistically correlated with the probability of fractures, which are an important public health problem, especially in elderly women. Therefore, BMD is an important indicator in diagnoses of osteoporosis, where patients might benefit from early management to improve their bone strength. For each sample, we assign an indicator based on three categories (normal, osteopenia, osteoporosis) before analyzing the data. The resulting hierarchical models with a pooling prior for BMD data outperformed the other models. To compare the models' performances, we calculated the DIC and the LCPO. Here, we found the best performance in the global–local pooling model. Although the nonparametric versions of the models have an infinite-dimensional parameter space, they showed similar values for the two comparison measures to those of the parametric pooling model with a finite-dimensional parameter space. Therefore, we should be careful in interpreting the results.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (2016R1D1A1B03932261). This research was also supported by a grant from the Simons Foundation (#353953, Balgobin Nandram).

Appendix

A. Computations for parameteric models

Let $\mathbf{n} = (\mathbf{n}_1, \ldots, \mathbf{n}_I)'$ be the response matrix and $\boldsymbol{\pi} = (\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_I)'$ be the proportion parameter matrix for (2.1). In an adaptive pooling model, let $\boldsymbol{\Omega}_i = (\boldsymbol{\pi}_i, \boldsymbol{\mu}, \tau)$, $i = 1, \ldots, I$. Here, no pooling and complete pooling are special cases of adaptive pooling, with parameters $\boldsymbol{\mu}$ and τ. The full conditional posterior density of the parameters for the given data is obtained in the usual way by combining the likelihood and the priors, as follows:

$$\begin{aligned}\pi(\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_I \,|\, \mathbf{n}) &\propto \left\{\prod_{i=1}^{I} f(\mathbf{n} \,|\, \boldsymbol{\pi}_i)\,\pi(\boldsymbol{\pi}_i)\right\}\pi(\boldsymbol{\mu}, \tau) \\ &\propto \prod_{i=1}^{I}\left\{\frac{1}{D(\boldsymbol{\mu}\tau)}\prod_{k=1}^{K}\pi_{ik}^{n_{ik}+\mu_k\tau-1}\right\}\frac{(K-1)!}{(\tau+1)^2},\end{aligned}$$

where $D(\boldsymbol{\mu}\tau) = \prod_{k=1}^{K}\Gamma(\mu_k\tau)\Big/\Gamma\left(\sum_{k=1}^{K}\mu_k\tau\right)$.

To run a Gibbs sampler, we draw values as follows:

(a) Full conditional for $\boldsymbol{\pi}_i$, $i = 1, \ldots, I$: Draw $\boldsymbol{\pi}_i \,|\, \mathbf{n}_i, \boldsymbol{\mu}, \tau \sim \text{Dirichlet}((\mathbf{n}_i + \boldsymbol{\mu})\tau)$.

(b) Full conditional for $\boldsymbol{\mu}$: Draw

$$\pi(\boldsymbol{\mu} \mid \mathbf{n}, \boldsymbol{\pi}, \tau) \propto \prod_{i=1}^{I} \left\{ \frac{\Gamma\left(\sum_{k=1}^{K} \mu_k \tau\right)}{\prod_{k=1}^{K} \Gamma(\mu_k \tau)} \prod_{k=1}^{K} \pi_{ik}^{\mu_k \tau - 1} \right\}. \tag{A.1}$$

Let $\boldsymbol{\mu}_{(k)}$, $k = 1, \ldots, K$ denote the vector of parameters other than the k^{th} component μ_k. Then, we obtain the conditional posterior density of μ_k, given $\boldsymbol{\mu}_{(k)}$, in each stage. Here, we need to estimate the $K - 1$ components of parameter $\boldsymbol{\mu}$ sequentially. Then, we can calculate the K^{th} component value of $\boldsymbol{\mu}$ using $\mu_K = 1 - \sum_{k=1}^{K-1} \mu_k$. Using the conditional posterior density (3.1), we can draw μ_k, $k = 1, \ldots, K - 1$ using the grid method, with support $1 - \sum_{k'=1,\, k' \neq k}^{K-1} \mu_{k'}$.

(c) Full conditional for τ: Draw

$$\pi(\tau \mid \mathbf{n}, \boldsymbol{\pi}, \boldsymbol{\mu}) \propto \prod_{i=1}^{I} \left\{ \frac{\Gamma\left(\sum_{k=1}^{K} \mu_k \tau\right)}{\prod_{k=1}^{K} \Gamma(\mu_k \tau)} \prod_{k=1}^{K} \pi_{ik}^{\mu_k \tau - 1} \right\} \frac{1}{(\tau + 1)^2}.$$

We can use the grid method for τ in this case as well. Because the grid method can be used for closed support, we transform τ to $\rho = 1/(1 + \tau)$, $0 < \rho < 1$. The absolute Jacobian is $1/\rho^2$. Then, the conditional posterior density of ρ can be expressed as follows:

$$\pi(\rho \mid \mathbf{n}, \boldsymbol{\pi}, \boldsymbol{\mu}) \propto \prod_{i=1}^{I} \left\{ \frac{\Gamma\left(\sum_{k=1}^{K} \mu_k \frac{1-\rho}{\rho}\right)}{\prod_{k=1}^{K} \Gamma\left(\mu_k \frac{1-\rho}{\rho}\right)} \prod_{k=1}^{K} \pi_{ik}^{\mu_k \frac{1-\rho}{\rho} - 1} \right\}.$$

The full conditional posterior density in the case of restricted pooling is obtained from the likelihood and the priors, constructed from $\pi(\boldsymbol{\mu}, \tau)$, and from an additional prior for ϕ, $i = 1, \ldots, I$:

$$\pi(\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_I \mid \mathbf{n}) = \left\{ \prod_{i=1}^{I} f(\mathbf{n} \mid \boldsymbol{\pi}_i)\, \pi(\boldsymbol{\pi}_i) \right\} \pi(\boldsymbol{\mu}, \tau)\, \pi(\phi),$$

where $\boldsymbol{\Omega}_i = (\boldsymbol{\pi}_i, \boldsymbol{\mu}, \tau, \phi)$, $i = 1, \ldots, I$.

For the Gibbs sampler, we consider the latent variables z_i, $i = 1, \ldots, I$ from a Bernoulli distribution with parameter ϕ. Here, the joint posterior density is given as follows:

$$\begin{aligned} \pi(\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_I \mid \mathbf{n}) &\propto \left\{ \prod_{i=1}^{I} f(\mathbf{n} \mid \boldsymbol{\pi}_i)\, \pi(\boldsymbol{\pi}_i \mid z_i)\, \pi(z_i) \right\} \pi(\boldsymbol{\mu}, \tau)\, \pi(\phi) \\ &\propto \left\{ \prod_{i=1}^{I} \left[\phi \frac{1}{D(\boldsymbol{\mu}\tau)} \prod_{k=1}^{K} \pi_{ik}^{\mu_k \tau + n_{ik} - 1} \right]^{z_i} \left[(1 - \phi) \prod_{k=1}^{K} \pi_{ik}^{n_{ik}} \right]^{1 - z_i} \right\} \\ &\quad \times \frac{(K-1)!}{(\tau + 1)^2} I_{\left(\phi \in \left(\frac{1}{2}, 1\right)\right)}. \end{aligned}$$

Then, our Gibbs sampler is described as follows:

(a) Full conditional for ϕ: Draw

$$\pi(\phi \mid \mathbf{n}, \boldsymbol{\mu}, \tau, \mathbf{z}) \propto \phi^{\sum_{i=1}^{I} z_i} (1 - \phi)^{I - \sum_{i=1}^{I} z_i}, \frac{1}{2} \leq \phi \leq 1.$$

In other words, given $\boldsymbol{\mu}$, τ, $\mathbf{z}$, and the data, ϕ follows a truncated beta distribution with parameters $\sum_{i=1}^{I} z_i$ and $I - \sum_{i=1}^{I} z_i$, and the lower bound of ϕ is $1/2$.

(b) Full conditional for z_i, $i = 1, \ldots, I$: Draw $z_i \,|\, \mathbf{n}, \boldsymbol{\pi}_i, \boldsymbol{\mu}, \tau \sim$ Bernoulli(p_i), where

$$p_i = \left(\phi \prod_{k=1}^{K} \pi_{ik}^{\mu_k \tau - 1} \Big/ D(\boldsymbol{\mu}\tau)\right) \Big/ \left(\phi \prod_{k=1}^{K} \pi_{ik}^{\mu_k \tau - 1} \Big/ D(\boldsymbol{\mu}\tau) + (1-\phi) I_{\left(\sum_{k=1}^{K} \pi_{ik} = 1\right)} \Big/ K\right).$$

(c) Full conditional for $\boldsymbol{\pi}_i$, $i = 1, \ldots, I$: Draw

$$\pi(\boldsymbol{\pi}_i \,|\, \mathbf{n}, z_i, \boldsymbol{\mu}, \tau) \propto \left[\prod_{k=1}^{K} \pi_{ik}^{\mu_k \tau + n_{ik} - 1}\right]^{z_i} \left[\prod_{k=1}^{K} \pi_{ik}^{1 + n_{ik} - 1}\right]^{1 - z_i}.$$

In the case of $z_i = 1$, we can generate the value of $\boldsymbol{\pi}_i$ from a Dirichlet distribution with parameters $\boldsymbol{\mu}$ and τ. In other cases, we can interpret that as the uncertainty of the modeling. That is, for $\boldsymbol{\pi}_i$ given $z_1 = 0$, $\mathbf{n}$ draws its value from a uniform Dirichlet distribution with a $K \times 1$ parameter vector, where each component has the value one.

(d) Full conditional for $\boldsymbol{\mu}$: Draw

$$\pi(\boldsymbol{\mu} \,|\, \mathbf{n}, \boldsymbol{\pi}, \tau) \propto \prod_{z_i = 1} \left\{\frac{\Gamma\left(\sum_{k=1}^{K} \mu_k \tau\right)}{\prod_{k=1}^{K} \Gamma(\mu_k \tau)} \prod_{k=1}^{K} \pi_{ik}^{\mu_k \tau - 1}\right\}.$$

(e) Full conditional for τ: Draw

$$\pi(\tau \,|\, \mathbf{n}, \boldsymbol{\pi}, \boldsymbol{\mu}) \propto \prod_{z_i = 1} \left\{\frac{\Gamma\left(\sum_{k=1}^{K} \mu_k \tau\right)}{\prod_{k=1}^{K} \Gamma(\mu_k \tau)} \prod_{k=1}^{K} \pi_{ik}^{\mu_k \tau - 1}\right\} \frac{1}{(\tau + 1)^2}.$$

Of course, the generating process for parameters $\boldsymbol{\mu}$ and τ is similar to that of the parameter in adaptive pooling. In addition, the data used for $\boldsymbol{\mu}$ and τ are $z_i = 1$, $i = 1, \ldots, I$.

Otherwise, the parameter vector corresponding to area i in global-local pooling consists of θ, $\boldsymbol{\eta}$, $\mathbf{z}$, ϕ, σ^2, σ_1^2, …, and σ_{K-1}^2. Then, the full conditional posterior density for the given data in the model is as follows:

$$\begin{aligned}
\pi(\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_I \,|\, \mathbf{n}) &\propto \left\{\prod_{i=1}^{I} f(\mathbf{n}_i \,|\, \boldsymbol{\eta}_i, \theta)\right\} \left\{\prod_{i=1}^{I} \pi(\boldsymbol{\eta}_i \,|\, z_i, \sigma^2, \sigma_1^2, \ldots, \sigma_{K-1}^2)\right\} \pi(\sigma^2) \\
&\quad \times \pi(\sigma_1^2, \ldots, \sigma_{K-1}^2)\, \pi(\theta)\, \pi(\mathbf{z} \,|\, \phi)\, \pi(\phi) \\
&= \left\{\prod_{i=1}^{I} \frac{n_i!}{\prod_{k=1}^{K} n_{ik}!} \left\{\prod_{k=1}^{K-1} \left(\frac{e^{\theta + \eta_{ik}}}{1 + \sum_{l=1}^{K-1} e^{\theta + \eta_{il}}}\right)^{n_{ik}}\right\} \left(\frac{1}{1 + \sum_{l=1}^{K-1} e^{\theta + \eta_{ik}}}\right)^{n_{iK}}\right\} \\
&\quad \times \left\{\prod_{i=1}^{I} \left[\prod_{k=1}^{K-1} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} \eta_{ik}^2\right)\right]^{z_i} \left[\prod_{k=1}^{K-1} \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{1}{2\sigma_k^2} \eta_{ik}^2\right)\right]^{1 - z_i}\right\} \\
&\quad \times \frac{1}{(1 + \sigma^2)^2} \left\{\prod_{k=1}^{K-1} \frac{1}{(1 + \sigma_k^2)^2}\right\} \frac{1}{\pi(1 + \theta^2)} \left\{\prod_{i=1}^{I} \phi^{z_i} (1 - \phi)^{1 - z_i}\right\} I_{(1/2 \le \phi \le 1)},
\end{aligned}$$

where $\boldsymbol{\Omega}_i = (\theta, \boldsymbol{\eta}_i, z_i, \phi, \sigma^2, \sigma_1^2, \ldots, \sigma_{K-1}^2)$, $i = 1, \ldots, I$, $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_I)'$ and $\mathbf{z} = (z_1, \ldots, z_I)'$.

Then, the Gibbs sampler is as follows:

(a) Full conditional for ϕ: Draw $\phi \,|\, \text{others} \sim \text{truncated Beta}\left(\sum_{i=1}^{I} z_i,\ I - \sum_{i=1}^{I} z_i,\ \frac{1}{2},\ 1\right)$.

(b) Full conditional for z_i, $i = 1, \ldots, I$: Draw $z_i \,|\, \text{others} \sim \text{Bernoulli}(p_i)$, with $p_i = \phi \prod_{k=1}^{K-1} \left(\exp(-\eta_{ik}^2 / 2\sigma^2)/\sqrt{2\pi\sigma^2}\right) \Big/ \left\{\phi \prod_{k=1}^{K-1} \exp(-\eta_{ik}^2/2\sigma^2)\Big/\sqrt{2\pi\sigma^2} + (1-\phi) \prod_{k=1}^{K-1} \left(\exp(-\eta_{ik}^2/2\sigma_k^2)/\sqrt{2\pi\sigma_k^2}\right)\right\}$.

(c) Full conditional for θ: Draw

$$\pi(\theta \,|\, \text{others}) \ \propto\ \left[\prod_{i=1}^{I}\prod_{k=1}^{K-1}\left(\frac{e^{\theta+\eta_{ik}}}{1+\sum_{l=1}^{K-1} e^{\theta+\eta_{il}}}\right)^{n_{ik}}\left(\frac{1}{1+\sum_{l=1}^{K-1} e^{\theta+\eta_{il}}}\right)^{n_{iK}}\right]\frac{1}{1+\theta^2}.$$

(d) Full conditional for η_{ik}, $i = 1, \ldots, I$, $k = 1, \ldots, (K-1)$: Draw

$$\pi(\eta_{ik} \,|\, \text{others}) \ \propto\ I_{(z_i=1)}\left(\frac{1}{1+\sum_{l=1}^{K-1} e^{\theta+\eta_{il}}}\right)^{\sum_{k=1}^{K} n_{ik}} \exp\left\{\sum_{k=1}^{K-1}\eta_{ik}n_{ik} - \frac{1}{2}\sum_{k=1}^{K-1}\frac{1}{\sigma^2}\eta_{ik}^2\right\}$$
$$+\ I_{(z_i=0)}\left(\frac{1}{1+\sum_{l=1}^{K-1} e^{\theta+\eta_{il}}}\right)^{\sum_{k=1}^{K} n_{ik}} \exp\left\{\sum_{k=1}^{K-1}\eta_{ik}n_{ik} - \frac{1}{2}\sum_{k=1}^{K-1}\frac{1}{\sigma_k^2}\eta_{ik}^2\right\}.$$

(e) Full conditional for σ^2: Draw

$$\pi(\sigma^2 \,|\, \text{others}) \ \propto\ \frac{1}{\sigma^{(K-1)\sum_{i=1}^{I} I(z_i=1)}\left(1+\sigma^2\right)^2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{z_i=1}\sum_{k=1}^{K-1}\eta_{ik}^2\right\}.$$

(f) Full conditional for σ_k^2 $k = 1, \ldots, (K-1)$: Draw

$$\pi(\sigma^2 \,|\, \text{others}) \ \propto\ \frac{1}{\sigma_k^{\sum_{i=1}^{I} I_{(z_i=0)}}\left(1+\sigma_k^2\right)^2} \exp\left\{-\frac{1}{2\sigma_k^2}\sum_{z_i=0}\eta_{ik}^2\right\}.$$

In this model, we suggest using the Metropolis-Hastings algorithm, which is the most commonly used Markov Chain Monte Carlo (MCMC) algorithm used to estimate the value of the location parameter, $(\theta, \boldsymbol{\eta})$. Of course, $(\sigma^2, \sigma_1^2, \ldots, \sigma_{(K-1)}^2)$ is drawn from the above full conditionals, and the Gibbs sampler is performed using the grid method.

B. Computations for nonparameteric models

In order to pool the parameters in nonparametric Bayesian models, we apply the slice sampling mehtod introduced by the Kalli, Griffin and Walker (2011). They proposed an efficient version of the slice sampler for Dirichlet process mixture models constructed by Walker (2007). Suppose that the observations y_i, $i = 1, \ldots, I$ are generated in the Dirichlet process mixture model with parameter θ. That is,

$$y_i \,|\, \mathrm{G} \overset{\text{iid}}{\sim} \mathrm{G}$$
$$\mathrm{G} \sim \mathrm{DP}(\alpha, \mathrm{G}_0)$$

Here, we write $\mathrm{G} \sim \mathrm{DP}(\alpha, \mathrm{G}_0)$ to denote that G follows a Dirichlet process with parameter $\alpha > 0$. Then G has a stick-breaking representation (Sethuraman, 1994) given by

$$P(y_i \,|\, G, \alpha) = \sum_{j=1}^{\infty} w_j f(y_i \,|\, \theta_j)$$

where $\theta_1, \theta_2, \theta_3, \ldots$ are independent and identically distributed (iid) form P_0 and

$$w_1 = \nu_1, \qquad w_j = \nu_j \prod_{l<j} (1 - \nu_l)$$

with the ν_j being iid from $\mathrm{Beta}(1, \alpha)$, see also Antoniak (1974).

The Slice sampler algorithm proposed by Walker (2007) introduces a latent variable $u \in (0, 1)$, $d_i = 1, 2, \ldots$ to perform sampling on the joint distribution. First, the latent variable u_i has a joint density as follow

$$P(y_i, u_i \,|\, G, \alpha) = \sum_{j=1}^{\infty} I_{(u_i < w_j)} f(y_i \,|\, \theta_j).$$

Later, they introduced a latent variable d_i representing the group assignment of the observation i. At this time, the joint density of (y_i, u_i, d_i) is as follow

$$P(y_i, u_i, d_i \,|\, G, \alpha) = \sum_{j=1}^{\infty} I_{(u_i < w_{d_i = j})} f(y_i \,|\, \theta_j).$$

Then we need to sample the parameter θ, α, and ν including latent variables u and d at each iteration of a Gibbs sampler. Kalli et al. (2011) introduces how to perform slice sampling for Dirichlet process mixture models by processing u and ν as blocks in the basic algorithm described by Walker (2007). The algorithm is as follow.

1. $\pi(\theta_j \,|\ldots) \propto G_0(\theta_j) \prod_{d_i = j} f(y_i \,|\, \theta_j)$,
2. $\pi(\nu_j) \propto \mathrm{Beta}(a_j, b_j)$, where $a_j = 1 + \sum_{i=1}^{I} I_{(d_i = j)}$ and $b_j = \alpha + \sum_{i=1}^{I} I_{(d_i > j)}$,
3. $\pi(u_i \,|\ldots) \propto I_{(0 < u_i < \xi_{d_i})}$, where $\xi_j = (1 - \kappa)\kappa^{j-1}$, and κ is a constant,
4. $P(d_i = k \,|\ldots) \propto I_{(k: \xi_k > u_i)} w_k / \xi_k f(y_i \,|\, \theta_k)$,
5. $\pi(\alpha \,|\ldots) \propto \alpha^J \prod_{j=1}^{J} (1 - \nu_j)^{\alpha - 1} \pi(\alpha)$.

For this paper, we need to sample the following variables at each iteration of a Gibbs sampler:

$$\{(\boldsymbol{\pi}_j, \nu_j),\ j = 1, 2, \ldots; (d_i, u_i),\ i = 1, \ldots, I\}.$$

In general, κ is equal to 0.5. However, we use κ as a tuning parameter for the hierarchical Bayesian model. Then, the full posterior density for the nonparametric adaptive pooling of he given data is given as follows:

$$P(\mathbf{n}, \mathbf{d}, \mathbf{u} \mid \boldsymbol{\nu}, \boldsymbol{\pi}) \left\{ \prod_{i=1}^{I} \pi(\boldsymbol{\pi}_i \mid \boldsymbol{\mu}, \tau) \right\} \pi(\boldsymbol{\mu}, \tau)\, \pi(\boldsymbol{\nu} \mid \alpha)\, \pi(\alpha)$$

$$\propto \left\{ \prod_{i=1}^{I} I_{(u_i < \xi_{d_i})} \frac{w_{d_i}}{\xi_{d_i}} n_{i\cdot}! \prod_{k=1}^{K} \frac{\pi_{d_i k}^{n_{ik}}}{n_{ik}!} \right\} \left\{ \prod_{i=1}^{I} \frac{1}{D(\boldsymbol{\mu}\tau)} \prod_{k=1}^{K} \pi_{ik}^{\mu_k \tau - 1} \right\}$$

$$\times \left\{ \frac{(K-1)!}{(1+\tau)^2} \right\} \left\{ \prod_{j=1}^{J} \frac{1}{B(1, \alpha)} (1 - \nu_j)^{\alpha - 1} \right\} \frac{1}{(1+\alpha)^2},$$

where $\mathbf{d} = (d_1, \ldots, d_I)$, $\mathbf{u} = (u_1, \ldots, u_I)$, $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_J)$, and the hyperparameters are mutually independent. Then, our Gibbs sampler can be performed in two steps.

The first step is the pooling of the data:

(a) Draw for u_i from Uniform$(0, \xi_{d_i})$.

(b) Draw for d_i from $P(d_i = j \mid \text{others}) = I_{(u_i < \xi_j)} w_j \Big/ \xi_j \prod_{k=1}^{K} \pi_{jk}^{n_{ik}} \Big/ n_{ik}!$. Next, we can generate the value of each parameter from the following conditional density:

(c) Draw $\boldsymbol{\pi}_j$, $j = 1, \ldots, J$, from Dirichlet$\left(\boldsymbol{\mu}\tau + \sum_{d_i = j} \mathbf{n}_i - 1\right)$.

(d) Draw ν_j, $j = 1, \ldots, J$, from Beta$\left(1 + \sum_{i=1}^{I} I_{(d_i = j)}, \alpha + \sum_{i=1}^{I} I_{(d_i = j)}\right)$.

(e) Draw $\boldsymbol{\mu}$ from

$$\pi(\boldsymbol{\mu} \mid \text{others}) \propto \prod_{j=1}^{J} \frac{1}{D(\boldsymbol{\mu}\tau)} \prod_{k=1}^{K} \pi_{jk}^{\mu_k \tau - 1}.$$

(f) Draw τ from

$$\pi(\tau \mid \text{others}) \propto \left\{ \prod_{j=1}^{J} \frac{1}{D(\boldsymbol{\mu}\tau)} \prod_{k=1}^{K} \pi_{jk}^{\mu_k \tau - 1} \right\} \frac{(K-1)!}{(\tau+1)^2}.$$

For our Gibbs sampler, we need to transform τ to $\rho = 1/(1+\tau)$, $0 < \rho < 1$ because we need to use the grid method for τ, which is taken from the noninformative prior with variable support equal to $(0, \infty)$.

$$\pi(\rho \mid \text{others}) \propto \left\{ \prod_{j=1}^{J} \frac{1}{D\left(\boldsymbol{\mu} \frac{1-\rho}{\rho}\right)} \prod_{k=1}^{K} \pi_{jk}^{\mu_k \frac{1-\rho}{\rho} - 1} \right\} (K-1)!.$$

(g) Draw α from

$$\pi(\alpha \mid \text{others}) \propto \left\{ \prod_{j=1}^{J} (1 - \nu_j)^{\alpha - 1} \right\} \frac{\alpha^J}{(1+\alpha)^2}.$$

The parameter α is also taken from the noninformative prior with variable support equal to $(0, \infty)$, as in the case of τ. Therefore, we need to transform α to δ, with Jacobian $1/\delta^2$. Then, the conditional density for δ is given as follows:

$$\pi(\delta \mid \text{others}) \propto \left\{ \prod_{j=1}^{J} (1 - \nu_j)^{\frac{1-\delta}{\delta} - 1} \right\} \left(\frac{1-\delta}{\delta} \right)^J.$$

The nonparametric version for the restricted pooling has $\boldsymbol{\pi}_j = (\pi_{j1}, \ldots, \pi_{jK})$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$, $0 \le \mu_k \le 1$, $\sum_{k=1}^{K} \mu_k = 1$, $\mathbf{d} = (d_1, \ldots, d_I)$, and $d_i = j$ for $j = 1, 2, \ldots, J$. Then, we compose the full posterior density for the given data using equation, as follows:

$$P(\mathbf{n},\, \mathbf{d},\, \mathbf{u}\,|\,\boldsymbol{\nu},\, \boldsymbol{\pi})\left\{\prod_{j=1}^{J}\pi\left(\boldsymbol{\pi}_j\,|\boldsymbol{\mu},\, \tau,\, \phi\right)\right\}\pi(\boldsymbol{\mu},\, \tau)\;\pi(\phi)\;\pi(\boldsymbol{\nu}\,|\alpha)\;\pi(\alpha).$$

In our Gibbs sampler, we consider the latent variables z_j, for $j = 1, \ldots, J$, as the parametric version for restricted pooling, where the subscripts for the parameter are equal to the parameter $\boldsymbol{\pi}$. At this time, the pooling step for the data is the same as above, and generating the parameter is as follows:

(a) Draw ϕ from truncated $\text{Beta}\left(\sum_{j=1}^{J} z_j,\; J - \sum_{j=1}^{J} z_j,\; \phi \in \left(\frac{1}{2}, 1\right)\right)$.

(b) Draw z_j from $\text{Bernoulli}\left(p_j\right)$,

where $p_j = \left(\phi\prod_{k=1}^{K}\pi_{jk}^{\mu_k\tau-1}\Big/D(\boldsymbol{\mu}\tau)\right)\Big/\left\{\phi\prod_{k=1}^{K}\pi_{jk}^{\mu_k\tau-1}\Big/D(\boldsymbol{\mu}\tau) + (1-\phi)\,I_{\left(\sum_{k=1}^{K}\pi_{jk}=1\right)}\Big/K-1\right\}$.

(c) Draw $\boldsymbol{\pi}_j$ from $\text{Dirichlet}\left(I_{(z_j=1)}\boldsymbol{\mu}\tau + I_{(z_j=0)}\;\mathbf{1} + \sum_{d_i=j}\mathbf{n}_i\right)$.

(d) Draw ν_j from $\text{Beta}\left(1 + \sum_{i=1}^{I} I_{(d_i=j)},\; \alpha + \sum_{i=1}^{I} I_{(d_i>j)}\right)$.

(e) Draw $\boldsymbol{\mu}$ from

$$\pi(\boldsymbol{\mu}\,|\,\text{others}) \;\propto\; \prod_{j=1}^{J}\frac{1}{D(\boldsymbol{\mu}\tau)}\prod_{k=1}^{K}\pi_{jk}^{\mu_k\tau-1}.$$

(f) Draw τ from

$$\pi(\tau\,|\,\text{others}) \;\propto\; \left[\prod_{j=1}^{J}\frac{1}{D(\boldsymbol{\mu}\tau)}\prod_{k=1}^{K}\pi_{jk}^{*\mu_k\tau-1}\right]\frac{(K-1)!}{(\tau+1)^2}.$$

(g) Draw α from

$$\pi(\alpha\,|\,\text{others}) \;\propto\; \left\{\prod_{j=1}^{m}(1-\nu_j)^{\alpha-1}\right\}\frac{\alpha^J}{(1+\alpha)^2}.$$

Lastly, the full posterior density is calculated using the joint posterior density and the priors for their parameters in the nonparametric Bayesian global-local pooling model. Here, the data pooling algorithm is the same as above inference. Then, we estimate each parameter as follows:

(a) Draw ϕ from truncated $\text{Beta}\left(\sum_{j=1}^{j} z_j,\; I - \sum_{j=1}^{J} z_j,\; \frac{1}{2},\; 1\right)$.

(b) Draw z_j from $\text{Bernoulli}\left(p_j\right)$, where $p_j = \phi\prod_{k=1}^{K-1}\left(\exp\left(-\eta_{jk}^2/2\sigma^2\right)\Big/\sqrt{2\pi\sigma^2}\right)\Big/\left\{\phi\prod_{k=1}^{K-1}\left(\exp(-\eta_{jk}^2/2\sigma^2)\Big/\sqrt{2\pi\sigma^2}\right) + (1-\phi)\prod_{k=1}^{K-1}\left(\exp\left(-\eta_{jk}^2/2\sigma_k^2\right)\Big/\sqrt{2\pi\sigma_k^2}\right)\right\}$.

(c) Draw θ from

$$\pi(\theta\,|\,\text{others}) \;\propto\; \left[\prod_{i=1}^{I}\prod_{k=1}^{K-1}\left(\frac{e^{\theta+\eta_{d_ik}}}{1+\sum_{l=1}^{K-1}e^{\theta+\eta_{d_il}}}\right)^{n_{ik}}\left(\frac{1}{1+\sum_{l=1}^{K-1}e^{\theta+\eta_{d_il}}}\right)^{n_{iK}}\right].$$

(d) Draw for η_{jk}, $k = 1, \ldots, (K-1)$:

$$\pi\left(\eta_{jk}\,|\,\text{others}\right) \;\propto\; I_{(z_j=1)}\left(\frac{1}{1+\sum_{l=1}^{K-1}e^{\theta+\eta_{jl}}}\right)^{\sum_{d_i=j}\sum_{k=1}^{K}n_{ik}}\exp\left\{\sum_{d_i=j}\sum_{k=1}^{K-1}\eta_{jk}n_{ik} - \frac{1}{2\sigma^2}\sum_{k=1}^{K-1}\eta_{jk}^2\right\}$$

$$+\; I_{(z_i=0)}\left(\frac{1}{1+\sum_{l=1}^{K-1}e^{\theta+\eta_{jl}}}\right)^{\sum_{d_i=j}\sum_{k=1}^{K}n_{ik}}\exp\left\{\sum_{d_i=j}\sum_{k=1}^{K-1}\eta_{jk}n_{ik} - \frac{1}{2}\sum_{k=1}^{K-1}\frac{1}{\sigma_k^2}\eta_{jk}^2\right\}.$$

(e) Draw ν_j from $\text{Beta}\left(1 + \sum_{i=1}^{I} I_{(d_i = j)},\ I_{(z_j=1)}\alpha + I_{(z_j=0)}\alpha_0 + \sum_{i=1}^{I} I_{(d_i > j)}\right)$.

(f) Draw σ^2 from

$$\pi(\sigma^2 \mid \text{others}) \propto \frac{1}{\sigma^{K-1\sum_{j=1}^{J} I_{(z_j=1)}} (1+\sigma^2)^2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{z_j=1}\sum_{k=1}^{K-1}\eta_{jk}^2\right\}.$$

(g) Draw σ_k^2 $k = 1, \ldots, (K-1)$ from

$$\pi(\sigma_k^2 \mid \text{others}) \propto \frac{1}{\sigma_k^{\sum_{j=1}^{J} I_{(z_j=0)}} (1+\sigma_k^2)^2} \exp\left\{-\frac{1}{2\sigma_k^2}\sum_{z_j=0}\eta_{jk}^2\right\}.$$

(h) Draw α from

$$\pi(\alpha \mid \text{others}) \propto \frac{\alpha^{\sum_{j=1}^{J} I_{(z_j=1)}}}{(1+\alpha)^2}\left(\prod_{z_j=1}(1-\nu_j)\right)^{\alpha-1}.$$

(i) Draw α_0 from

$$\pi(\alpha_0 \mid \text{others}) \propto \frac{\alpha_0^{\sum_{j=1}^{J} I_{(z_j=0)}}}{(1+\alpha_0)^2}\left(\prod_{z_j=0}(1-\nu_j)\right)^{\alpha_0-1}.$$

References

Agresti, A., and Hitchcock, D.B. (2005). Bayesian inference for categorical data analysis. *Statistical Methods and Applications*, 14, 297-330.

Antoniak, C.E. (1974). Mixture of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2, 1152-1174.

Consonni, G., and Veronese, P. (1995). A Bayesian method for combining results from several binomial experiments. *Journal of the American Statistical Association*, 90, 935-944.

DuMouchel, W.H., and Harris, J.E. (1983). Bayes methods for combining the results of cancer studies in humans and other species. *Journal of the American Statistical Association*, 78, 293-308.

Dunson, D.B. (2009). Nonparametric Bayes local partition models for random effects. *Biometrika*, 96, 249-262.

Evans, R., and Sedransk, J. (1999). Methodology for pooling subpopulation regressions when sample sizes are small and there is uncertainty about which subpopulations are similar. *Statistica Sinica*, 9, 345-359.

Evans, R., and Sedransk, J. (2003). Bayesian methodology for combining the results from different experiments when the specifications for pooling are uncertain: II. *Journal of Statiatical Planning and Inference*, 111, 95-100.

Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1, 209-230.

Gneiting, T., and Reftery, A.E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 359-378.

Ishwaran, H., and James, L.F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96, 161-173.

Kalli, M., Griffin, J.E. and Walker, S.G. (2011). Slice sampling mixture models. *Statistics and Computing*, 21, 93-105.

Leonard, T. (1977). Bayes simultaneous estimation for several multinomial distributions. *Communications in Statistics: Theory and Methods*, 6, 619-630.

Malec, D., and Sedransk, J. (1992). Bayesian methodology for combining the results from different experiments when the specifications for pooling are uncertain. *Biometrika*, 79, 593-601.

Nandram, B. (1998). A Bayesian analysis of the three-stage hierarchical multinomial model. *Journal of Statistical Computation and Simulation*, 61, 97-112.

Nandram, B., and Yin, J. (2016a). Bayesian predictive inference under a Dirichlet process with sensitivity to the normal baseline. *Statistical Methodology*, 28, 1-17.

Nandram, B., and Yin, J. (2016b). A nonparametric Bayesian prediction interval for a finite population mean. *Journal of Statistical Computation and Simulation*, 86, 3141-3157.

Nandram, B., Zhou, J. and Kim, D.H. (2019). A pooled Bayes test of independence for sparse contingency tables from small area. *Journal of Statistical Computation and Simulation*, 89(5), 899-926.

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

Sethuraman, J. (1994). A constructive definition of dirichlet priors. *Statistica Sinica*, 4, 639-650.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and Linde, A.V.D. (2002). Bayesian measures of model complexity and fit. *Journal of royal statistical society, Series B*, 64(4), 538-639.

Walker, S.G. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics-Simulation and Computation*, 36, 45-54.

Yin, J., and Nandram, B. (2020a). A nonparametric Bayesian analysis of response data with gaps, outlier and ties. *Statistics and Applications*, 18(2), 121-141.

Yin, J., and Nandram, B. (2020b). A Bayesian small area model with Dirichlet processes on the responses. *Statistics in Transition*, 21(1), 1-17.

Survey Methodology, June 2021
Vol. 47, No. 1, pp. 215-222
Statistics Canada, Catalogue No. 12-001-X

A note on multiply robust predictive mean matching imputation with complex survey data

Sixia Chen, David Haziza and Alexander Stubblefield[1]

Abstract

Predictive mean matching is a commonly used imputation procedure for addressing the problem of item nonresponse in surveys. The customary approach relies upon the specification of a single outcome regression model. In this note, we propose a novel predictive mean matching procedure that allows the user to specify multiple outcome regression models. The resulting estimator is multiply robust in the sense that it remains consistent if one of the specified outcome regression models is correctly specified. The results from a simulation study suggest that the proposed method performs well in terms of bias and efficiency.

Key Words: Multiple robustness; Nearest-neigbour imputation; Survey data; Variance estimation.

1. Introduction

Predictive mean matching (PMM), a procedure closely related to nearest-neighbour imputation (NNI, Chen and Shao, 2000; Beaumont and Bocci, 2009; Yang and Kim, 2019), is a popular imputation procedure in practice (Little, 1988; Yang and Kim, 2020). In NNI, a missing value to a survey variable y is replaced by the y-value of the closest respondent with respect to a vector of fully observed variables $\mathbf{x}$. However, with NNI, the resulting imputed estimator may suffer from a non-negligible bias when the dimension of $\mathbf{x}$ is large (Yang and Kim, 2019), a problem often referred to as the curse of dimensionality. In contrast, PMM starts with fitting a parametric model (e.g., a linear regression model) based on the responding units with y as the response variable and $\mathbf{x}$ as the set of explanatory variables. This leads to a set of predicted values or scores, $\hat{m}$, for all the sample units (respondents and nonrespondents). A missing value to the survey variable y is then replaced by the y-value of the closest respondent with respect to $\hat{m}$. The latter may be viewed as a scalar summary of the information contained in the vector $\mathbf{x}$. Therefore, unlike NNI, PMM is not sensitive to the dimension of $\mathbf{x}$, which is a desirable feature.

Both NNI and PMM belong to the class of nonparametric procedures. Therefore, both procedures are less vulnerable to model misspecification unlike parametric methods (e.g., linear regression imputation). Also, both NNI and PMM belong to the class of donor imputation procedures; that is, they produce eligible imputed values as they use actual observed values "borrowed" from the respondents.

In the first step of PMM, the information contained in the vector $\mathbf{x}$ is compressed into a single score $\hat{m}$ through the use of a parametric model (e.g., a linear regression model). If the specified model provides an accurate description of the relationship linking y and $\mathbf{x}$, we expect PMM to perform well in terms of bias. On the other hand, if the specified model is grossly misspecified, PMM may yield biased estimators.

1. Sixia Chen, Department of Biostatistics and Epidemiology, University of Oklahoma Health Sciences Center, Oklahoma City, OK 73104, U.S.A.; David Haziza, Department of mathematics and statistics, University of Ottawa, Ottawa, Canada. E-mail: dhaziza@uottawa.ca; Alexander Stubblefield, Department of Economics, Michigan State University, East Lansing, MI 48823, U.S.A.

Multiply robust approaches with multiple outcome regression and nonresponse models have been shown to improve the robustness against model misspecification, see Han and Wang (2013), Han (2014), and Chen and Haziza (2019a) among others. In this note, we propose a novel PMM procedure that allows for multiple models, each which may be based on a different functional and/or a different set of explanatory variables. Postulating multiple models may prove useful in a number of situations; e.g., see Chen and Haziza (2017) and Chen and Haziza (2019b) for a discussion. The specified models may be parametric or nonparametric. The rationale behind the proposed method is to fit each of these specified models based on the responding units, which leads to multiple set of predicted values (scores) for all the sample units. After describing the theoretical setup in Section 2, we show how to combine these scores to construct the imputed values in Section 3. The proposed PMM procedure is multiply robust in the sense that the resulting estimator is consistent if all but one model are misspecified. Because the true model linking y and $\mathbf{x}$ is unknown, the proposed approach is attractive because it provides some protection against model misspecification. Also, unlike the multiply robust imputation procedure considered in Chen and Haziza (2017), the proposed method belongs to the class of donor imputation procedures. In Section 4, we conduct a simulation study to assess the performance of the proposed method in terms of bias and efficiency.

2. Basic setup

Consider a finite population $\mathcal{F}_N = \{(\mathbf{x}_i, y_i),\ i = 1, 2, \ldots, N\}$, assumed to have been generated from the following superpopulation model:

$$y_i = m(\mathbf{x}_i) + \varepsilon_i, \tag{2.1}$$

where $m(\cdot)$ is an unknown functional, $\mathbf{x}_i$ is a vector of fully observed variables attached to unit i, and the ε_i's are mutually independent random variables such that $\mathbb{E}(\varepsilon_i | \mathbf{x}_i) = 0$ and $\mathbb{V}(\varepsilon_i | \mathbf{x}_i) = \sigma^2$. For simplicity, we assume that the variance structure is homoscedastic but our method can be easily extended to the case of unequal variances.

The interest lies in estimating the population mean, $\theta = \mathbb{E}(y)$. Given the finite population, a probability sample S, of size n, is selected according to a sampling design with first-order inclusion probabilities π_i and second-order inclusion probabilities π_{ij}. The sampling weight attached to unit i is denoted by $w_i = \pi_i^{-1}$.

Let r_i be response indicator attached to unit i such that $r_i = 1$ if y_i is observed, and $r_i = 0$ if y_i is missing. Let $S_r = \{i \in S: r_i = 1\}$ denote the set of respondents to the survey variable y. We assume that the data are Missing At Random (MAR):

$$\Pr(r_i = 1 | \mathbf{x}_i, y_i) = \Pr(r_i = 1 | \mathbf{x}_i). \tag{2.2}$$

The customary PMM procedure can be described as follows. We first postulate a parametric outcome regression model $\mathcal{M} = \{m(\mathbf{x}_i; \boldsymbol{\beta})\}$, where $\boldsymbol{\beta}$ is a vector of unknown parameters (Yang and Kim, 2020). For $i \in S$, we compute the score $\hat{m}_i = m(\mathbf{x}_i; \hat{\boldsymbol{\beta}})$, where $\hat{\boldsymbol{\beta}}$ is a suitable estimator of $\boldsymbol{\beta}$ based on the

responding units. Then, the imputed value for the missing y_i is $y_i^* = y_j$, where j is the index of the nearest-neighbour of unit i, which satisfies $\mathcal{D}(\hat{m}_j, \hat{m}_i) \leq \mathcal{D}(\hat{m}_{j'}, \hat{m}_i)$ for any $j' \in S_r$, where $\mathcal{D}(\cdot,\cdot)$ denotes a distance function; e.g., the Euclidean distance. In order for PMM to be robust against misspecification, the specified parametric model must satisfy the Lipschitz continuity condition (Yang and Kim, 2020). This condition may not be satisfied for some commonly used models and functional forms, including quadratic models; see Yang and Kim (2020) for a discussion.

3. Proposed method

The proposed method allows the user to specify multiple outcome regression models for the survey variable y. This grants a greater probability of selecting a model that performs well at replicating the relationship between the response variable and the explanatory variables, making the approach multiply robust without requiring the Lipschitz continuity condition to hold. As long as one of the specified models is correctly specified, the resulting estimator will be consistent.

We consider a class of outcome regression models: $\mathcal{M} = \left\{m^{(k)}\left(\mathbf{x}_i; \boldsymbol{\beta}^{(k)}\right), k = 1, 2, \ldots, K\right\}$. To impute the missing values, we proceed as follows:

(Step1). Obtain the estimators $\hat{\boldsymbol{\beta}}^{(k)}$ of $\boldsymbol{\beta}^{(k)}$, $k = 1, 2, \ldots, K$, by solving the following survey weighted estimating equations:

$$\hat{U}_m^{(k)}\left(\hat{\boldsymbol{\beta}}^{(k)}\right) = \sum_{i \in S} w_i r_i \left\{y_i - m^{(k)}\left(\mathbf{x}_i; \boldsymbol{\beta}^{(k)}\right)\right\} \frac{\partial m^{(k)}\left(\mathbf{x}_i; \boldsymbol{\beta}^{(k)}\right)}{\partial \boldsymbol{\beta}^{(k)}} = 0. \tag{3.1}$$

(Step2). For $i \in S$, obtain the K-vector of predicted values

$$\mathbf{V}_i = \left(m^{(1)}\left(\mathbf{x}_i; \hat{\boldsymbol{\beta}}^{(1)}\right), m^{(2)}\left(\mathbf{x}_i; \hat{\boldsymbol{\beta}}^{(2)}\right), \ldots, m^{(K)}\left(\mathbf{x}_i; \hat{\boldsymbol{\beta}}^{(K)}\right)\right)^{\top}.$$

(Step3). Fit a weighted linear regression model without intercept with y as the response variable and $\mathbf{V}$ as the vector of explanatory variables. Let $\hat{M}\left(\mathbf{x}_i; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\eta}}\right)$ be the resulting predicted value attached to unit i:

$$\hat{M}\left(\mathbf{x}_i; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\eta}}\right) = \mathbf{V}_i^{\top} \hat{\boldsymbol{\eta}},$$

where $\hat{\boldsymbol{\beta}} = \left(\hat{\boldsymbol{\beta}}^{(1)}, \hat{\boldsymbol{\beta}}^{(2)}, \ldots, \hat{\boldsymbol{\beta}}^{(K)}\right)$ and

$$\hat{\boldsymbol{\eta}} = \left\{\sum_{i \in S} w_i r_i \mathbf{V}_i \mathbf{V}_i^{\top}\right\}^{-1} \sum_{i \in S} w_i r_i \mathbf{V}_i y_i. \tag{3.2}$$

(Step4). The imputed value for the missing y_i is $y_i^* = y_j$, where j is the index of the nearest-neighbour of unit i, which satisfies $\mathcal{D}\{\hat{M}(\mathbf{x}_j; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\eta}}), \hat{M}(\mathbf{x}_i; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\eta}})\} \leq \mathcal{D}\{\hat{M}(\mathbf{x}_{j'}; \hat{\boldsymbol{\eta}}), \hat{M}(\mathbf{x}_i; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\eta}})\}$ for any $j' \in S_r$.

After applying (Step1)-(Step4), we construct the imputed estimator of θ:

$$\hat{\theta}_{\mathrm{MR}} = \frac{1}{\hat{N}} \sum_{i \in S} w_i \left\{ r_i y_i + (1 - r_i)\, y_i^* \right\}, \tag{3.3}$$

where $\hat{N} = \sum_{i \in S} w_i$. Using an approach similar to the one used by Yang and Kim (2020), it can be shown that the estimator $\hat{\theta}_{\mathrm{MR}}$ is multiply robust in the sense that it is consistent if all but one model are misspecified.

Estimating the variance of $\hat{\theta}_{\mathrm{MR}}$ can be done through replication variance estimation procedures; see e.g., Rust and Rao (1996) and Wolter (2007). In the context of PMM for survey data, Yang and Kim (2020) also considered replication procedures. Let L denote the number of replicates and $w_i^{(g)}$ be a replication weight attached to unit i in the g^{th} replicate. A replication variance estimator of $\hat{\theta}_{\mathrm{MR}}$ is given by

$$\hat{V}_{\mathrm{rep}}\left(\hat{\theta}_{\mathrm{MR}}\right) = \sum_{g=1}^{L} c_g \left(\hat{\theta}_{\mathrm{MR}}^{(g)} - \hat{\theta}_{\mathrm{MR}}\right)^2, \tag{3.4}$$

where

$$\hat{\theta}_{\mathrm{MR}}^{(g)} = \sum_{i \in S} w_i^{(g)} \left\{ r_i y_i + (1 - r_i)\, y_i^{*(g)} \right\}$$

denote the estimator $\hat{\theta}_{\mathrm{MR}}$ in the g^{th} replicate with $y_i^{*(g)}$ denoting the imputed value attached to unit i in the g^{th} replicate, obtained from (Step1)-(Step4) above, based on the replication weight $w_i^{(g)}$ instead of the original weights w_i. The factor c_g in (3.4) is determined by the replication method. For instance, with the delete-one jackknife, we have $L = n$, $c_g = n/(n-1)$ and $w_i^{(g)} = n/(n-1)\, w_i$ if $i \neq g$ and $w_i^{(g)} = 0$ if $i = g$.

4. Simulation study

To assess the performance of the proposed method in terms of bias and efficiency, we conducted a limited simulation study. We generated $B = 2{,}000$ finite populations, each of size $N = 20{,}000$. First, the explanatory variables x_1-x_4 were generated from a multivariate standard normal distribution. Then, given x_1-x_4, we generated the survey variable y according to the following outcome regression models:

(M1). $y = 1 + x_1 + x_2 + x_3 + x_4 + \varepsilon$, where $\varepsilon \sim N(0, 1)$.

(M2). $y = 1 + x_1^2 + x_2^2 + x_3 + x_4 + x_3 x_4 + \varepsilon$, where $\varepsilon \sim N(0, 1)$.

Note that both (M1) and (M2) are linear models based on the explanatory variables x_1-x_4, except that (M2) includes quadratic terms and an interaction term.

From each finite population, a probability sample S was selected according to probability proportional-to-size (PPS) systematic sampling based on the size variable $z_i = \log(0.1 \mid y_i + v_i \mid + 4)$,

where $v_i \sim N(0, 1)$. The first-order inclusion probabilities are given by $\pi_i = nz_i \Big/ \sum_{i=1}^{N} z_i$ with n = 200, 500 and 1,000.

In each sample, the response indicators r_i were generated from a Bernoulli distribution with probability p_i, where

$$p_i = 0.1 + 0.9 \times \frac{\exp(\alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_3 x_{3i} + \alpha_4 x_{4i})}{1 + \exp(\alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_3 x_{3i} + \alpha_4 x_{4i})}. \tag{4.1}$$

We used two sets of values for $(\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4)$: $(0, 1, 1, 1, 1)$ and $(1.38, 1, 1, 1, 1)$. These led to response rates approximately equal to 70%, and 50%, respectively.

We computed the following estimators of θ

(Naive). The weighted mean of the respondents, $\hat{\theta}_{\text{naive}} = \sum_{i \in S_r} w_i y_i \Big/ \sum_{i \in S_r} w_i$.

(Reg). The imputed estimator based on deterministic linear regression imputation, assuming the model (M1).

(PMM1). The imputed estimator based on PMM, where the score $\hat{m}_i$, $i \in S$, was obtained by fitting the model (M1).

(New1). The imputed estimator based on the proposed multiply robust PMM procedure using both models (M1) and (M2).

(New2). The imputed estimator based on the proposed multiply robust PMM procedure using models (M1), (M2), and two additional models (M3) and (M4), where (M3) uses x_1 only as the predictor and (M4) uses x_1^2 only as the predictor.

We computed the Monte Carlo relative bias (MCRB), the Monte Carlo relative standard error (MCRSE) and the Monte Carlo relative root mean squared error (MCRMSE), defined respectively as

$$\text{MCRB} = \frac{2{,}000^{-1} \sum_{b=1}^{2{,}000} (\hat{\theta}_b - \theta_b)}{\theta_{\text{MC}}},$$

$$\text{MCRSE} = \frac{\sqrt{(B-1)^{-1} \sum_{b=1}^{B} (\hat{\theta}_b - \hat{\theta}_{\text{MC}})^2}}{\theta_{\text{MC}}}$$

and

$$\text{MCRMSE} = \frac{\sqrt{(B-1)^{-1} \sum_{b=1}^{B} (\hat{\theta}_b - \theta_{\text{MC}})^2}}{\theta_{\text{MC}}},$$

where θ_b denotes the population mean in the b^{th} population, $\hat{\theta}_b$ denotes the estimator $\hat{\theta}$ in the b^{th} sample, b = 1, …, 2,000, and

$$\theta_{\text{MC}} = \frac{1}{2{,}000} \sum_{b=1}^{2{,}000} \theta_b, \quad \hat{\theta}_{\text{MC}} = \frac{1}{2{,}000} \sum_{b=1}^{2{,}000} \hat{\theta}_b.$$

The results are presented in Tables 4.1 and 4.2. The naive estimator exhibited a significant bias in all the scenarios, as expected. When the true model was given by (M1), we note from Table 4.1 that linear regression imputation performed very well in terms of bias, as expected. Both PMM and the proposed method showed negligible bias for n = 1,000 and a slight bias for n = 500 and n = 200. For instance, for n = 200 and a response rate of 70%, the value of RB was equal to 2.4% for PMM, New1 and New2. In terms of efficiency, linear regression imputation slightly outperformed both PMM and the proposed methods, as expected. For instance for n = 1,000 and a response rate of 70%, the value of RMSE was equal to 7.5% for linear regression imputation and equal to 8.0% for both PMM, New1 and New2. It is worth pointing out that both PMM and the proposed methods exhibited almost identical performances in all the scenarios presented in Table 4.1. Therefore, incorporating two additional models did not seem to affect the efficiency of the resulting estimator (New2).

When the true model was given by (M2), we note from Table 4.2 that both linear regression imputation and PMM led to significant biases in all the scenarios, as expected. Being a parametric imputation procedure, linear regression imputation is vulnerable to model misspecification. On the other hand, PMM showed smaller biases than linear regression imputation, suggesting some robustness against model misspecification. For instance, for n = 1,000 and a response rate of 70%, the value of RB was equal to -9.2% for linear regression imputation and -3.7% for PMM. The proposed methods outperformed both linear regression imputation and PMM in terms of bias, standard error and mean square error in all the scenarios. Finally, both New1 and New2 exhibited almost identical performances.

Table 4.1
Monte Carlo relative bias (MCRB), relative standard error (MCRSE), and relative root mean squared error (MCRMSE) when the true model is (M1)

			Method				
Response rate	**Sample Size**	**Measure $(\times 10^2)$**	**Naive**	**Reg**	**PMM1**	**New1**	**New2**
70%	1,000	MCRB	64.7	-0.1	0.4	0.4	0.4
		MCRSE	7.5	7.5	8.0	8.0	8.0
		MCRMSE	65.1	7.5	8.0	8.0	8.0
70%	500	MCRB	65.3	0.5	1.4	1.4	1.4
		MCRSE	10.7	10.4	11.2	11.2	11.2
		MCRMSE	66.1	10.4	11.3	11.3	11.3
70%	200	MCRB	64.6	0.3	2.4	2.4	2.4
		MCRSE	16.5	16.7	17.5	17.5	17.6
		MCRMSE	66.7	16.7	17.7	17.7	17.7
50%	1,000	MCRB	99.3	0.0	0.7	0.7	0.6
		MCRSE	8.8	8.1	9.0	9.0	9.0
		MCRMSE	99.7	8.1	9.1	9.1	9.1
50%	500	MCRB	98.9	-0.1	1.3	1.3	1.3
		MCRSE	12.1	11.2	12.5	12.5	12.5
		MCRMSE	99.6	11.2	12.6	12.6	12.6
50%	200	MCRB	99.8	0.8	4.3	4.3	4.4
		MCRSE	19.3	17.7	19.6	19.6	19.6
		MCRMSE	101.6	17.7	20.1	20.1	20.0

Table 4.2
Monte Carlo relative bias (MCRB), relative standard error (MCRSE), and relative root mean squared error (MCRMSE) when the true model is (M2)

			Method				
Response rate	**Sample Size**	**Measure $(\times 10^2)$**	**Naive**	**Reg**	**PMM1**	**New1**	**New2**
70%	1,000	MCRB	7.5	-9.2	-3.7	0.1	0.1
		MCRSE	3.5	3.5	3.9	3.1	3.1
		MCRMSE	8.2	9.9	5.4	3.1	3.1
70%	500	MCRB	7.5	-9.4	-4.0	0.2	0.2
		MCRSE	5.0	5.1	5.6	4.5	4.5
		MCRMSE	9.0	10.7	6.9	4.5	4.5
70%	200	MCRB	7.6	-9.2	-4.0	0.1	0.1
		MCRSE	7.8	7.9	8.5	6.8	6.8
		MCRMSE	10.9	12.1	9.4	6.8	6.8
50%	1,000	MCRB	16.6	-11.3	-3.1	0.3	0.3
		MCRSE	4.0	4.5	5.0	3.3	3.3
		MCRMSE	17.1	12.2	5.9	3.3	3.3
50%	500	MCRB	16.5	-11.5	-3.5	0.3	0.3
		MCRSE	5.7	6.3	7.0	4.8	4.7
		MCRMSE	17.5	13.2	7.8	4.8	4.8
50%	200	MCRB	16.5	-12.0	-3.9	-0.1	-0.1
		MCRSE	9.1	9.9	11.0	7.4	7.4
		MCRMSE	18.8	15.6	11.7	7.4	7.4

Acknowledgements

S. Chen was supported by the National Institute on Minority Health and Health Disparities (NIMHD) at National Institutes of Health (NIH) (1R21MD014658-01A1) and the Oklahoma Shared Clinical and Translational Resources (U54GM104938) with an Institutional Development Award (IDeA) from National Institute of General Medical Sciences. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The work of D. Haziza was supported by grants from the Natural Sciences and Engineering Research Council of Canada.

References

Beaumont, J.-F., and Bocci, C. (2009). Variance estimation when donor imputation is used to fill in missing values. *Canadian Journal of Statistics*, 37, 400-416.

Chen, S., and Haziza, D. (2017). Multiply robust imputation procedures for the treatment of item nonresponse in surveys. *Biometrika*, 102, 439-453.

Chen, S., and Haziza, D. (2019a). Multiply robust nonparametric multiple imputation for the treatment of missing data. *Statistica Sinica*, 29, 2035-2053.

Chen, S., and Haziza, D. (2019b). Recent developments in dealing with item nonresponse in surveys: A critical review. *International Statistical Review*, 87, S192-S218.

Chen, J., and Shao, J. (2000). Nearest-neighbour imputation for survey data. *Journal of Official Statistics*, 16, 583-599.

Han, P. (2014). Multiply robust estimation in regression analysis with missing data. *Journal of the American Statistical Association*, 109, 1159-1173.

Han, P., and Wang, L. (2013). Estimation with missing data: Beyond double robustness. *Biometrika*, 100, 417-430.

Little, R.J.A. (1988). Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics*, 6, 287-296.

Rust, K.F., and Rao, J.N.K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, 5, 283-310.

Wolter, K. (2007). *Introduction to Variance Estimation, 2nd Edition*. Springer, Berlin.

Yang, S., and Kim, J.K. (2019). Nearest neighbor imputation for general parameter estimation in survey sampling. *Advances in Econometrics - The Econometrics of Complex Survey Data: Theory and Applications*, 39, 209-234.

Yang, S., and Kim, J.K. (2020). Predictive mean matching imputation in survey sampling. To appear in the *Scandinavian Journal of Statistics*.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 36, No. 4, December 2020

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 37, No. 1, March 2021

All inquires about submissions and subscriptions should be directed to jos@scb.se

CONTENTS TABLE DES MATIÈRES

Volume 48, No. 3, September/septembre 2020

Original Articles

CONTENTS TABLE DES MATIÈRES

Volume 48, No. 4, December/décembre 2020

Original Articles

GUIDELINES FOR MANUSCRIPTS

Authors are invited to submit their articles **through the *Survey Methodology* hub on the ScholarOne Manuscripts website** (https://mc04.manuscriptcentral.com/surveymeth). Before submitting the article, please examine a recent issue of *Survey Methodology* 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.